Post selection shrinkage estimation for high-dimensional data analysis

By: Xiaoli Gao, S. E. Ahmed, Yang Feng

This is the peer reviewed version of the following article:

Gao, X.L., Ahmed, S.E. and Feng, Y. (2017). (Discussion paper) Post Selection Shrinkage Estimation for High Dimensional Data Analysis, *Applied Stochastic Models in Business and Industry, 33(2), 97-120.* doi:10.1002/asmb.2193,

which has been published in final form at http://dx.doi.org/10.1002/asmb.2193 . This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Abstract:

In high-dimensional data settings where $p \gg n$, many penalized regularization approaches were studied for simultaneous variable selection and estimation. However, with the existence of covariates with weak effect, many existing variable selection methods, including Lasso and its generations, cannot distinguish covariates with weak and no contribution. Thus, prediction based on a subset model of selected covariates only can be inefficient. In this paper, we propose a post selection shrinkage estimation strategy to improve the prediction performance of a selected subset model. Such a post selection shrinkage estimator (PSE) is data adaptive and constructed by shrinking a post selection weighted ridge estimator in the direction of a selected candidate subset. Under an asymptotic distributional quadratic risk criterion, its prediction performance is explored analytically. We show that the proposed post selection PSE performs better than the post selection weighted ridge estimator. More importantly, it improves the prediction performance of any candidate subset model selected from most existing Lasso-type variable selection methods significantly. The relative performance of the post selection PSE is demonstrated by both simulation studies and real-data analysis.

Keywords: asymptotic risk | lasso | ridge regression | (positive) shrinkage estimation | post selection | sparse model

Articles:

1 Introduction

Many high-dimensional data arise in biological, medical, social, and economical studies. Because of the trade-off between model complexity and model prediction, the statistical inference of model selection becomes extremely important and challenging in high-dimensional data analysis. Consider a classical high-dimensional linear regression model with *i*th observed response variable *yi* and covariates *xij*s,

$$y_i = \sum_{j=1}^{p_n} x_{ij}\beta_j + \varepsilon_i, \quad 1 \leqslant i \leqslant n, \tag{1.1}$$

where ϵis is independent and identically distributed random errors with center 0 and variance σ^2. Without loss of generality, we do not include the intercept in the model by assuming all data have been centered. Here, the subscript *n* in *pn* indicates that the number of coefficients may increase with the sample size *n*. Such a notation will be used throughout the paper without further explanation.

Over the past two decades, many penalized regularization approaches have been developed to do variable selection and estimation simultaneously. Among them, the Lasso [1] is one of the most popular approaches because of its convexity and computation efficiency. In general, the Lasso penalty tends to select an over-fitted model because it penalizes all coefficients equally [2]. Many endeavors have been undertaken to improve the Lasso to reach both variable selection consistency and the estimation consistency. To list a few, smoothly clipped absolute deviation [3, 4], adaptive Lasso [5] and minimax concave penalty [6], among others. An overview of variable selection in high-dimensional feature space can be found in [7].

In order to have nice estimation and selection properties, most Lasso-type penalties make some important assumptions about both true model and designed covariates. For example, the true model is often assumed to be sparse, insofar that (i) most βjs are zeros except for a few ones and (ii) all those nonzero βj's are larger than an inflated noise level, $c\sigma\sqrt{(2/n)\log(p_n)}$ with $c \geqslant 1/2$ [8]. Additional assumptions made on the designed covariates include the adaptive irrepresentable condition and the restricted eigenvalue conditions. For detailed information, we refer to [9], [10], and [11].

However, those conditions are somewhat restrictive and are not judiciously justified in real applications. Consequently, Lasso and its generalizations may have lower prediction efficiency once those assumptions are violated. To fix the idea, we take the sparse model assumption (ii) as an example. Suppose we can divide the index set $\{1,\ldots,pn\}$ into three disjoint subsets: S_1, S_2, and S_3. In particular, S_1includes indexes of nonzero βi's which are moderately large and easily detected; S_3 includes indexes with only zero coefficients; S_2, being the intermediate, includes indexes of those nonzero βj with weak but nonzero effects. Thus, S_1 is able to be detected using some existing variable selection techniques, while S_2 may not be separated from S_3 in general using existing Lasso-type methods. A more detailed description can be found in [8]. Following the spirit of model parsimony, covariates in S_1 are kept in the model, and some or all covariates in S_2 are left aside with ones in S_3. Author in [12] has showed using simulation studies that such a Lasso estimate often performs worse than the post selection least squares estimate. To improve the prediction error of a Lasso-type variable selection approach, some (modified) post least squares estimators are studied in [13] and [14]. However, this work still assume the irrepresentable condition, and those post estimations are only based upon the chosen subset after the Lasso. Consequently, the simultaneous weak effects in S_2 are still ignored. An ideal strategy would be able to incorporate the joint contribution from covariates in S_2, even though a parsimonious model without including covariates in S_2 is adopted.

Let us consider an extreme case where S_1 is a null set and p is fixed. It has been studied extensively that shrinkage estimators can have uniformly smaller risk compared with the ordinary maximum likelihood estimators (MLEs) since the discussion papers in [15] and [16]. The relative risk properties of shrinkage estimators were also investigated in low-dimensional regression models under a restricted linear submodel space. See, for example, [17-20] and many others.

However, in high-dimensional settings where $p > n$, *a priori* information on S_1 is not guaranteed, not mentioning the existence of an MLE. Thanks to the existing variable selection techniques, an estimated candidate subset $\hat{S}_1$ is selected. Once $\hat{S}_1$ is obtained, the next question we want to ask is: can we construct a post selection shrinkage estimate to improve the risk of the post selection least squares estimators?

As we know, ridge regression [21, 22] has been widely used when the design matrix is ill-conditioned such that a regular MLE is not available. In this paper, we follow the model parsimony spirit and extend shrinkage estimation to the high-dimensional data setting using both ridge penalty and Lasso-type penalty separately. In particular, we use a ridge penalty to construct a data-adaptive post selection shrinkage estimator (PSE) to improve the risk of a post selection least squares estimator based upon a Lasso-type variable selection result.

We summarize our main contributions as follows:

1. We propose a post selection shrinkage strategy to improve the risk of the Lasso-type estimators in high-dimensional settings. This post selection shrinkage strategy is data adaptive and has some practical applications, especially when an 'important' subset is generated and some covariates with joint weak effects are not selected.

2. We investigate the asymptotic risk of the proposed PSEs. Corresponding asymptotic properties of a predecessor generating those PSEs are also investigated under some regularity conditions.

The rest of the paper is organized as follows. In Section 2, we describe some preliminary model information involved in building a PSE. As preparation, we introduce some sparsity definitions under certain signal strength levels. Some existing variable selection results from Lasso are also summarized in this section. We propose three steps in constructing the shrinkage strategy in Section 3. In Section 4, we investigate some asymptotic properties of those post selection estimators during three steps in Section 3. We first investigate some asymptotic normality properties of the designed weighted ridge (WR) estimators under some conditions. Then, we investigate the asymptotic distributional risks of the linear combination of the proposed PSEs. In Sections 5 and 6, we perform some numerical studies using some simulated examples and a real-data application, respectively. We summarize the paper with some discussions in the final section. All proofs are given in the Appendix.

2 Model description and basic notations

Let $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_{p_n}^*)'$ be the true coefficie ts vector in model (1.1). For any subset $S \subset \{1,\ldots,pn\}$ with a cardinal value $|S|$, denote $\boldsymbol{\beta}_S^*$ a subvector of $\boldsymbol{\beta}^*$ indexed by S. Similar subscripts are used for other submatrices and subvectors.

2.1 Model sparsity and signal strength

As introduced in the previous section, the effect of all pn covariates is characterized into three categories based upon their signal strength: important covariates with strong effects in S_1, covariates with no effect in S_3, and an intermediate group in S_2 with joint weak effects. In particular, those *signal strength assumptions* of the true model are made explicitly as follows:

(**A1**) There exists a positive constant c_1, such that $|\beta_j^*| > c_1\sqrt{(\log p_n)/n}$ for $\forall j \in S_1$;

(**A2**) The parameter vector $\boldsymbol{\beta}^*$ satisfies that $\|\boldsymbol{\beta}_{S_2}^*\| = O(n^\tau)$ for some $0 < \tau < 1$, where $\|\cdot\|$ is the ℓ_2 norm;

(**A3**) $\beta_j^* = 0$, for $\forall j \in S_3$.

Assumptions (A1–A3) specify those signal strength levels in the strong signals set S_1, weak signals set S_2, and sparse signal set S_3 explicitly. In particular, (A2) indicates that joint weak effects in $\boldsymbol{\beta}_{S_2}^*$ only grow with n at a certain rate, even though the dimension pn grows with n fast. For example, if (A1) holds for some $c_1 > 0$ and we let $|\beta_{0j}| < c_1\sqrt{(\log p_n)/n}$ for $j \in S_2$ with $|S_2| < n$, then $\|\boldsymbol{\beta}_{S_2}^*\| < c_1\sqrt{\log(p_n)} < O(n^\tau)$ even though $pn = O(\exp(n^2\tau))$.

Most existing high-dimensional sparse models investigate the variable selection consistency by only considering the existence of the strong signals in (A1) and sparse signals in (A3). There is very limited work assuming the existence of weak signals in S_2. For example, besides a strong signal set in (A1), [23] does not separate S_2 and S_3 and makes an alternative sparse model assumption,

(A2') $\sum_{j \notin S_1} |\beta_j^*| \leqslant \eta_1$ for some $\eta_1 \geqslant 0$.

In their work, some sufficient conditions are investigated under which the Lasso can select the strong signal set S_1 consistently, following the spirit of the model parsimony.

Our weak and sparse conditions in (A2–A3) are different from the sparse condition in (A2') where S_2 and S_3 are not separated. If we replace (A2) by (A2') in our signal strength assumptions, then (A2) becomes $\|\boldsymbol{\beta}_{S_2}^*\| \leqslant \sum_{j \in S_2} |\beta_j^*| = \eta_1$, the joint effects in S_2 being bounded uniformly. Thus, a true model under (A2') only is less sparse than one under (A3) only but more sparse than one in both (A2) and (A3). On the contrary, a sparse model under both (A2) and (A3) includes the most weak signals; a sparse model under (A3) only does not have any weak signals, while a sparse model under (A2') only is in the middle.

2.2 Parsimonious model selection

As discussed in Section 1, a penalized least squares (PLS) estimator is often adopted to select a parsimonious model for a high-dimensional regression model in (1.1),

$$\hat{\boldsymbol{\beta}}_n^{\mathrm{PLS}} = \arg\ \min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p_n} x_{ij}\beta_j \right)^2 + \sum_{j=1}^{p_n} p_\lambda(\beta_j) \right\}, \tag{2.1}$$

where $p\lambda(\beta j)$ is the penalty term on βj with a tuning parameter controlling the size of selected candidate subset model. For example, the Lasso takes $p\lambda(\beta j) = \lambda|\beta j|$, and the adaptive Lasso takes $p\lambda(\beta j) = \lambda|\beta j|/|wj|$, where wj can be taken as an initial estimator of βj. The size of selected subset model depends strongly on the choice of tuning parameters in (2.1). As pointed out by [8], one turns to ignore weak signals in S_2together with S_3 and select a candidate subset model with only strong signals in S_1, following the model parsimony spirit.

If we let $\hat{S}_1 \subset \{1, \ldots, p_n\}$ index an active subset from (2.1), then a data-adaptive candidate subset model is produced such that

$$\hat{\beta}_j^{\mathrm{PLS}} = 0 \quad \text{if and only if } j \notin \hat{S}_1. \tag{2.2}$$

Denote the response vector $\mathbf{y} = (y_1,\ldots,yn)'$, all covariates vectors $\mathbf{x}j = (x_1j,\ldots,xnj)'$ for $j = 1,\ldots,pn$, and the design matrix $\mathbf{X} = (\mathbf{x}_1 \ldots \mathbf{x}_{p_n})$. Without loss of generality, we rearrange the designed vectors such that $\mathbf{X} = (\mathbf{X}_{S_1}|\mathbf{X}_{S_2}|\mathbf{X}_{S_3})$, where $\mathbf{X}S$ is the submatrix consists of vectors indexed by $S\subset\{1,\ldots,pn\}$. Next we give two scenarios where S_2 cannot be separated from S_3.

Case 1. ([24]) Consider an orthonormal design with $\mathbf{X}'\mathbf{X}/n = \mathbf{I}n$ and $\boldsymbol{\epsilon} \sim N(0,\mathbf{I}n)$. The PLS with Lasso penalty provides a soft-threshold estimator with $\hat{\beta}_j^{\mathrm{Lasso}} = \tilde{\beta}_j - \lambda/(2n)\mathrm{sgn}(\tilde{\beta}_j)$ and 0, for $|\tilde{\beta}_j| > \lambda/(2n)$ and $|\tilde{\beta}_j| < \lambda/(2n)$, respectively. Here, $\tilde{\beta}_j = \mathbf{x}_j'\mathbf{y}/n \sim N(\beta_{0j}, 1/n)$ is the least squares solution, and sgn(·) is the sign mapping function.
If $\min_{j\in S_1} |\beta_j^*| > \lambda/(2n) > c > \max_{j\in S_2} |\beta_j^*|$ for some $c > 0$, then $P(\hat{S}_1 = S_1) \to 1$; that is, $P(\hat{\beta}_j^{\mathrm{PLS}} = 0) \to 1$ for $j \notin S_1$. Thus, all weak signals in S_2 are omitted together with sparse signals in S_3 using the Lasso approach.

Case 2. ([25]) Consider a non-singular design such that the smallest eigenvalue of $\mathbf{X}'_{S_3^c}\mathbf{X}_{S_3^c}/n$ is larger than some positive constant c. If there exists some $j\in S_2$ such that $|\beta_0 j| < |gj(\lambda)|$, where $g_j(\lambda) = \lambda\mathbf{e}_j'(\mathbf{X}'_{S_3^c}\mathbf{X}_{S_3^c})^{-1}\mathrm{sgn}(\boldsymbol{\beta}_{0S_3^c})$ with $\mathbf{e}j$ being the jth column of the identity matrix, then $P(\{S_1 \cup S_2 \subseteq \hat{S}_1\} \cap \{S_3 \subseteq \hat{S}_1^c\}) < 1$. Thus, S_2 and S_3 cannot be separated using the Lasso.

Some post selection estimators were proposed to improve the prediction performance of the PLS estimator. For example, under some regularity conditions, [13] and [14] studied some post selection least square estimators,

$$\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1} = \left(\mathbf{X}_{\hat{S}_1}{}'\mathbf{X}_{\hat{S}_1}\right)^{-1}\mathbf{X}_{\hat{S}_1}{}'\mathbf{y}. \quad (2.3)$$

Here, we denote such a post selection least squares estimator as a restricted estimator (RE), written as $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}$ in this paper. For notation's convenience, we omit the phase of 'post selection' in some future short notations without causing any confusion.

When S_1 and S_2 are not separable, we tend to select the important subset $\hat{S}_1$, such that $\hat{S}_1 \subseteq S_1$ for a large enough λ, or $S_1 \subset \hat{S}_1 \subset S_1 \cup S_2$ for a smaller λ, following the spirit of model parsimony. Although $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}$ is more estimation efficient than $\hat{\boldsymbol{\beta}}^{\mathrm{PLS}}_{\hat{S}_1}$, the prediction risk of $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}$ can still be high because many weak signals in S_2 are ignored in $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}$. Our interest is to improve the risk performance of $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}$ given in (2.3) by picking up some information from $\hat{S}_1^c$, a complement subset of the selected candidate submodel.

2.3 Some additional notations

Based upon a subset partition S_1,S_2,S_3, we can partition the true parameters $\boldsymbol{\beta}^* = \left(\boldsymbol{\beta}_1^{*\prime}, \boldsymbol{\beta}_2^{*\prime}, \boldsymbol{\beta}_3^{*\prime}\right)'$, without loss of generality. Some notations are shortened for notation's simplicity such that $\boldsymbol{\beta}^*_{S_k} = \boldsymbol{\beta}^*_k$ for k = 1,2 and 3. Similar notations are also adopted for other subvectors and matrices. For example, after the same partition, the design matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_{p_n})$ can be written as $\mathbf{X} = (\mathbf{X}_1,\mathbf{X}_2,\mathbf{X}_3)$. We also write $\mathbf{X} = (\mathbf{Z},\mathbf{X}_3)$ with $\mathbf{Z} = (\mathbf{X}_1,\mathbf{X}_2)$. The row vector of $\mathbf{Z}$ is denoted as $\mathbf{z}_i = (z_{i1}, \cdots, z_{i,p_1+p_2})$ for $1 \le i \le n$.

We denote $p_k = |S_k|$ for $1 \le k \le 3$ and $p_n = p_1 + p_2 + p_3$. In this paper, we allow $p_n = \sum_{k=1}^{3} p_k$ to be very large but restrict $q = p_1 + p_2 \le n$ such that $\boldsymbol{\Sigma}_n = n^{-1}\mathbf{Z}'\mathbf{Z}$ is non-singular. If $\boldsymbol{\Sigma}_n$ is singular, then a generalized inverse matrix is adopted when needed in computations. Some other submatrices of $\boldsymbol{\Sigma}_n$ are defined as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_{n11} &= \mathbf{X}_1{}'\mathbf{X}_1/n, \quad \boldsymbol{\Sigma}_{n22} = \mathbf{X}_2{}'\mathbf{X}_2/n, \\
\boldsymbol{\Sigma}_{n12} &= \mathbf{X}_1{}'\mathbf{X}_2/n, \quad \boldsymbol{\Sigma}_{n21} = \mathbf{X}_2{}'\mathbf{X}_1/n, \\
\boldsymbol{\Sigma}_{n22.1} &= n^{-1}\mathbf{X}_2{}'\mathbf{X}_2 - \mathbf{X}_2{}'\mathbf{X}_1(\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1{}'\mathbf{X}_2 \\
\boldsymbol{\Sigma}_{n11.2} &= n^{-1}\mathbf{X}_1{}'\mathbf{X}_1 - \mathbf{X}_1{}'\mathbf{X}_2(\mathbf{X}_2{}'\mathbf{X}_2)^{-1}\mathbf{X}_2{}'\mathbf{X}_1 \quad (2.4)
\end{aligned}$$

Let $\mathbf{U} = (\mathbf{X}_2,\mathbf{X}_3)$ be a $n \times (p_n - p_1)$ submatrix of $\mathbf{X}$. Then, another partition is written as $\mathbf{X} = (\mathbf{X}_1,\mathbf{U})$. Let $\boldsymbol{M}_1 = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$. Then, $\mathbf{U}'\boldsymbol{M}_1\mathbf{U}$ is a $(p_n - p_1) \times (p_n - p_1)$ dimensional singular matrix with rank $k_n \geqslant 0$. We denote $\varrho_{1n} \leqslant \ldots \leqslant \varrho_{k_n n}$ as all k_n positive eigenvalues of $\mathbf{U}'\boldsymbol{M}_1\mathbf{U}$.

3 Post selection shrinkage estimation strategy

We propose a high-dimensional post selection shrinkage estimation strategy based upon the following three steps:

Step 1: Obtain a data-adaptive candidate subset $\hat{S}_1$ following a model parsimony spirit and construct a post selection least square estimator $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}$ using (2.3);

Step 2: Obtain a post selection WR estimator, $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_n = \left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1}, \hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1^c}\right)$, using a threshold ridge penalty to be introduced and a submodel $\hat{S}_1$ selected from Step 1;

Step 3: Obtain a PSE by shrinking $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{\hat{S}_1}$ from Step 2 in the direction of $\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{\hat{S}_1}$ from Step 1.

The post selection WR estimator in Step 2 can handle three scenarios simultaneously: (a) the sparsity in high-dimensional data analysis; (b) the strong correlation among covariates; and (c) the jointly weak contribution from some covariates.

Remark 1. This post selection shrinkage estimation is expected to improve the risk performance on the selected submodel once a variable selection approach in Step 1 tends to select those and only those variables with strong signal strength, that is, $S_1 \supset \hat{S}_1$ or $S_1 \subset \hat{S}_1 \subset S_1 \cup S_2$. However, if the model parsimony spirit is not followed and λ in (2.1) is too small such that $\hat{S}_1 \supset S_1 \cup S_2$, this post selection shrinkage estimation is not suggested. Therefore, the effect of the PSE is data adaptive and depends on $\hat{S}_1$.

As a preparation, we first construct a post selection WR estimation based upon $\hat{S}_1$. This post selection weight ridge estimation itself is constructed from two steps introduced in Section 'Weighted ridge estimation' and 'Post selection shrinkage estimation'.

3.1 Weighted ridge estimation

Once $\hat{S}_1$ is obtained from Step 1, we seek to minimize a penalized objective function with a ridge penalty on coefficients in $\hat{S}_1^c$,

$$\widetilde{\boldsymbol{\beta}}(r_n) = \arg\ \min\{L(\boldsymbol{\beta}; \hat{S}_1)\} = \arg\ \min\left\{\|\mathbf{Y} - \mathbf{X}_n\boldsymbol{\beta}_n\|^2 + r_n\|\boldsymbol{\beta}_{\hat{S}_1^c}\|^2\right\} \quad (3.1)$$

where $rn > 0$ is a tuning parameter controlling the penalty effect on $\boldsymbol{\beta}_{\hat{S}_1}$. Then, a post selection WR estimator $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}(r_n, a_n; \hat{S}_1) = (\beta^{\mathrm{WR}}_{\hat{S}_1}(r_n), \beta^{\mathrm{WR}}_{\hat{S}_1^c}(r_n, a_n)$ is obtained from,

$$\hat{\beta}^{\mathrm{WR}}_j(r_n, a_n) = \begin{cases} \tilde{\beta}_j(r_n), & j \in \hat{S}_1; \\ \tilde{\beta}_j(r_n) I\left(\tilde{\beta}_j(r_n) > a_n\right), & j \in \hat{S}_1^c, \end{cases} \quad (3.2)$$

where $I(\cdot)$ is the indicator function and *an* is a threshold parameter. Thus, we obtain estimators of the weak signal subset

$$\hat{S}_2 := \hat{S}_2(\hat{S}_1) = \left\{ j \in \hat{S}_1^c : \hat{\beta}_j^{\mathrm{WR}}(r_n, a_n) \neq 0 \right\} \quad (3.3)$$

and of the sparse subset

$$\hat{S}_3 := \hat{S}_3(\hat{S}_1) = \left(\hat{S}_1 \cup \hat{S}_2 \right)^c. \quad (3.4)$$

Our post selection strategy is only applied when the threshold parameter *an* satisfies $|\hat{S}_2| > 2$ and $|\hat{S}_3^c| < n$. In particular, we set

$$a_n = c_1 n^{-\alpha}, \quad 0 < \alpha \leqslant 1/2, \text{ for some } c_1 > 0. \quad (3.5)$$

Remark 2. We call $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}(r_n, a_n)$ a post selection WR estimator from two facts: (i) we only penalize parameters in $\boldsymbol{\beta}_{\hat{S}_1^c}$ instead of the entire coefficients vector $\boldsymbol{\beta}n$, and (ii) the threshold step in (3.2) can be interpreted as a WR penalty $r_n \sum_{j \in \hat{S}_1^c} \left(\beta_j^2 / w_j^2 \right)$ in (3.1), where $wj = 0$ and 1 for $j \in \hat{S}_3$ and $j \in \hat{S}_2$.

Remark 3. Similar to the discussion in Remark 2, we can also understand the post selection step into the WR estimator, $r_n \sum_{j \in \hat{S}_1^c} \left(\beta_j^2 / w_j^2 \right)$ with $wj = \infty$ for $j \in \hat{S}_1$. We do not enforce an additional ridge penalty on $\hat{S}_1$ to reduce some unnecessary biases during the WR step. This is different from the post selection threshold regression studied in [26], where the ℓ_2 penalty is applied on the entire $\boldsymbol{\beta}n$ equally.

Remark 4. The idea of the WR regression is connected to the regularization after retention framework proposed in [27]. In that framework, a retention step is conducted to find the important set $\hat{S}_1$ with large marginal-correlation coefficients with the response. Then, a regularization step is conducted by a penalized least square with L_1 regularization only on the covariates that are not in $\hat{S}_1$. Compared with that framework, the current framework focused more on prediction by using the ridge penalty, and the estimate $\hat{S}_1$ is also different.

Notice that for every selected candidate subset $\hat{S}_1$, $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{WR}}(r_n)$ depends on *rn* and $\hat{\boldsymbol{\beta}}_{\hat{S}_1^c}^{\mathrm{WR}}(r_n, a_n)$ depends on both *rn* and *an*. For convenience, we omit those tuning parameters and denote above post selection WR estimators as $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{WR}}$ and $\hat{\boldsymbol{\beta}}_{\hat{S}_1^c}^{\mathrm{WR}}$, respectively.

3.2 Post selection shrinkage estimation

Now, we are ready to propose a shrinkage estimation based upon two post selection estimators: $\hat{\boldsymbol{\beta}}_{S_1}^{\mathrm{RE}}$ and $\hat{\boldsymbol{\beta}}_{S_1}^{\mathrm{WR}}$.

An initial PSE $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{SE}}$ is defined as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{SE}} &= \hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{RE}} + \left(\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{RE}}\right)\left(1 - (\hat{s}_2 - 2)/\hat{T}_n\right) \\ &= \hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{WR}} - \left((\hat{s}_2 - 2)/\hat{T}_n\right)\left(\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{RE}}\right), \end{aligned} \quad (3.6)$$

where $\hat{s}_2 = |\hat{S}_2|$ and $\hat{T}_n$ are given by

$$\hat{T}_n = \left(\hat{\boldsymbol{\beta}}_{\hat{S}_2}^{\mathrm{WR}}\right)' \left(\mathbf{X}_{\hat{S}_2}' \boldsymbol{M}_{\hat{S}_1} \mathbf{X}_{\hat{S}_2}\right) \hat{\boldsymbol{\beta}}_{\hat{S}_2}^{\mathrm{WR}} / \sigma^2, \quad (3.7)$$

where $\boldsymbol{M}_{\hat{S}_1} = \mathbf{I}_n - \mathbf{X}_{\hat{S}_1}\left(\mathbf{X}_{\hat{S}_1}' \mathbf{X}_{\hat{S}_1}\right)^{-1} \mathbf{X}_{\hat{S}_1}'$. If σ^2 is unknown, it is replaced by a consistent estimator $\hat{\sigma}^2$. In the numerical studies, σ^2 is replaced by $\hat{\sigma}^2 = \sum_{i=1}^{n}\left(y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\hat{S}_2}^{\mathrm{WR}}\right)^2 / (n - \hat{s}_2)$, and a generalized inverse is used if $\left(\mathbf{X}_{\hat{S}_1}' \mathbf{X}_{\hat{S}_1}\right)^{-1}$ is not singular.

Observing from (3.6) and (3.7), signs of two estimators of $\boldsymbol{\beta}_{\hat{S}_1}$ can be reversed if $\hat{T}_n$ is too small such that $\hat{s}_2 - 2 > \hat{T}_n$. It is possible because $\hat{\boldsymbol{\beta}}_{\hat{S}_1^c}^{\mathrm{WR}}$ consists of nuisance parameters, and over-shrinkage can occur for a large *rn* in the WR step. Thus, we also suggest to modify (3.6) as the following post selection PSE,

$$\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{PSE}} = \hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{WR}} - \left([(\hat{s}_2 - 2)/\hat{T}_n] \wedge 1\right)\left(\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{RE}}\right). \quad (3.8)$$

Remark 5. Our proposed post selection shrinkage estimation and the classical shrinkage estimation bear some resemblance but are different because of two facts: (i) Post selection shrinkage estimation is associated with a selected candidate subset and has some flexibility of adjusting the shrinkage strength data adaptively because $\hat{\boldsymbol{\beta}}_{\hat{S}_1^c}^{\mathrm{WR}}$ depends on tuning parameters *an* and *rn*; (ii) Post selection shrinkage estimation uses an initial ridge shrinkage step and is tailored for the high-dimensional settings where multiple covariates tend to be correlated and function jointly.

4 Asymptotic properties

In order to investigate some asymptotic properties of the proposed post selection estimators, we first make following assumptions on the random error, $\mathbf{U}'\boldsymbol{M}_1\mathbf{U}$, and the model sparsity. One can review some notations at the end of Section 2.

(B1) The random error $\epsilon i \sim N(0,\sigma^2)$.

(B2) $\varrho_{1n}^{-1} = O(n^{-\eta})$, where $\tau < \eta \leq 1$ for τ in (A2).

(B3) $\log(pn) = O(nv)$ for $0 < v < 1$.

(B4) There exists a positive definite matrix $\mathbf{\Sigma}$ such that $\lim_{n\to\infty} \mathbf{\Sigma}_n = \mathbf{\Sigma}$, where eigenvalues of $\mathbf{\Sigma}$ satisfy $0 < \rho_1 < \rho\mathbf{\Sigma} < \rho_2 < \infty$.

Here, condition (B1) can be relaxed to a symmetric distribution with some finite moments. To simplify our theoretical investigations and handle the ultra high dimensionality, we only restrict our studies to normal random error in this paper. Condition (B2) guarantees that the positive eigenvalues of the redundant $\boldsymbol{U}^{\prime}\boldsymbol{M}_1\boldsymbol{U}$ cannot be too small with a rate associated with the weak signals strength in S_2. Condition (B3) permits the ultra-high dimensionality such that the number of variables can grow with sample size at an almost exponential rate. Condition (B4) is the regularity condition for $\mathbf{X}_{S_3^c}$. This condition is made in order to obtain the asymptotic normality the WR estimator.

4.1 Asymptotic properties of the weighted ridge estimator

We have the following asymptotic properties of the WR estimator $\hat{\boldsymbol{\beta}}_n^{\mathrm{WR}}$.

Theorem 1. Suppose the sparse model in (1.1) satisfies signal strength assumptions in (A1–A3) and model assumptions in (B1–B3). If we choose $r_n = c_2 a_n^{-2} (\log\log n)^3 \log(n \vee p_n)$ for some constant $c_2 > 0$ and *an* defined in (3.5) with $\alpha < (\eta - v - \tau)/3$, then $\hat{S}_2$ in (3.3) satisfies

$$P\left(\hat{S}_2 = S_2 | \hat{S}_1 = S_1\right) \geqslant 1 - (n \vee p_n)^{-t} \text{ for some constant } t > 0, \tag{4.1}$$

where τ, η, and v are defined in (A2), (B2), and (B3), respectively.

Theorem 1 is similar to the variable selection result in [28]. We postpone the detailed proof to the Appendix. It tells us that the WR estimator $\hat{\boldsymbol{\beta}}_{S_1^c}^{\mathrm{WR}}$ is able to single out the sparse set S_3 with a large probability, if S_1 is pre-selected in advance such that $P(\hat{S}_1 = S_1) = 1$. For example, [23] argued that S_1 can be recovered with a large probability under the sparse Riesz condition (SRC) with rank p_1. Here, a design matrix $\mathbf{X}$ satisfies the SRC with rank q and spectrum bounds $0 < c_* < c^* < \infty$ if

$$c_* \leqslant \frac{\|\mathbf{X}_S \mathbf{v}\|^2}{\|\mathbf{v}\|^2} \leqslant c^* \quad \forall S \text{ with } |S| = q \text{ and } \mathbf{v} \in \mathcal{R}^q. \tag{4.2}$$

Lemma 1. Consider the Lasso solution for linear model (1.1) with $\epsilon i \sim N(0,\sigma^2)$. Suppose (A1) and (B1) are satisfied, and the sparse condition (A2') holds for some $0 < \eta_1 < O(p_1\sqrt{\log(p_n)/n})$,

and the design matrix **X** satisfies the SRC with rank p_1 in (4.2). Then, $\hat{S}_1$ generated from a PLS with the Lasso penalty in (2.1) satisfies

$$\lim_{n\to\infty} P\left(\{S_1 \subset \hat{S}_1\} \cap \left\{\sum_{j\in S_1} |\beta_j^*| I\left(\hat{\beta}_j^{\mathrm{PLS}} = 0\right) = 0\right\}\right) = \lim_{n\to\infty} P(\widehat{S_1} = S_1) = 1.$$

Lemma 1 is a direct result from Theorem 2 in [23]. Here, the tuning parameter in (2.1) is chosen such that $\lambda \geqslant 2\sigma\sqrt{2(1+c_0)c^* n \log(p_n)}$. Lemma 1 indicates that those and only those strong signals in S_1 are included in $\hat{S}_1$ while using the Lasso under sufficient conditions.

In Lemma 1, we have $\sum_{j\notin S_1} |\beta_j^*| < \eta$. The signal of each individual coefficient is trivial if such a joint effect is uniformly distributed on $pn - p_1$ coefficients when $pn \gg n$. However, if this joint effect is only distributed on a much smaller number of coefficients, each individual effect may not be negligible. In particular, if we let both (A2') and (A3) hold, then $\sum_{j\in S_2} |\beta_j^*| < \eta$. Thus, (A2) also holds. Combing Lemma 1 and Theorem 1, we have the following result directly.

Corollary 1. Suppose all conditions in both Lemma 1 and Theorem 1 hold. Then, we have

$$\lim_{n\to\infty} P\left(\{\hat{S}_2 = S_2\} \cap \{\hat{S}_1 = S_1\}\right) = 1. \tag{4.3}$$

Corollary 1 indicates that $\hat{S}_3 = S_3$ is able to be recovered if an additional WRs step is used post the Lasso under some sufficient conditions. We skip the proof because this is a direct result from Lemma 1 and Theorem 1.

However, Corollary 1 still requires a SRC condition. Although $P(\hat{S}_1 = S_1) = 1$ may not be guaranteed when a SRC condition is not satisfied, we may have

$$P\left(\{S_1 \subset \hat{S}_1 \subset S_1 \cup S_2\}\right) \to 1. \tag{4.4}$$

Thus, we have similar but weaker result.

Corollary 2. Suppose all conditions in Theorem 1 hold, and $\hat{S}_1$ satisfies (4.4). Then, we have

$$\lim_{n\to\infty} P\left(\{\hat{S}_2 = \hat{S}_1^c \cap S_2\}\right) = 1. \tag{4.5}$$

Corollary 2 can be interpreted by treating $\hat{S}_1$ as a new S_1 and $\hat{S}_1^c \cap S_2$ as a new S_2.

The asymptotic properties in Theorem 1 and its derivatives in Corollary 1 and 2 are important for establishing the efficiency of $\hat{\boldsymbol{\beta}}_{\hat{S}_1}^{\mathrm{WR}}$ and $\hat{\boldsymbol{\beta}}_{\hat{S}_2}^{\mathrm{WR}}$.

Theorem 2. Let $s_n^2 = \sigma^2 \mathbf{d}_n' \boldsymbol{\Sigma}_n^{-1} \mathbf{d}_n$ for any $(p_1 n + p_2 n) \times 1$ vector $\mathbf{d}n$ satisfying $\|\mathbf{d}n\| \leq 1$. Suppose assumptions (B1–B4) hold. Consider a sparse model with signal strength under (A1), (A3), and

(A2) with $0 < \tau < 1/2$. Suppose a pre-selected model such as $S_1 \subset \hat{S}_1 \subset S_1 \cup S_2$ is obtained with probability 1. If we choose *rn* as in Theorem 1 with $\alpha < \{(\eta - \nu - \tau)/3, 1/4 - \tau/2\}$, then we have the asymptotic normality,

$$n^{1/2} s_n^{-1} \mathbf{d}_n' \left(\hat{\boldsymbol{\beta}}_{S_3^c}^{\mathrm{WR}} - \boldsymbol{\beta}_{S_3^c}^* \right) \xrightarrow{\mathrm{d}} N(0, 1). \quad (4.6)$$

Theorem 2 studies the asymptotic normality of the WR estimator, $\hat{\boldsymbol{\beta}}_{S_3^c}$. In addition, $\hat{\boldsymbol{\beta}}_{S_3^c}$ has the same estimation efficiency as one from a restricted least square estimator as if $\boldsymbol{\beta}_{S_3} = 0$ is given as *a priori*. However, the result holds if $\|\boldsymbol{\beta}_{S_2}^*\| = o(n^{1/2})$ and *rn* is chosen appropriately. More importantly, the strong signal set S_1 is detected with a large probability in advance. This can be guaranteed under Lemma 1.

4.2 Asymptotic distributional risk analysis

In this section, we provide the relative performance of the post selection shrinkage estimation regarding the asymptotic distribution risk (ADR) introduced in [29]. For simplicity and notation's convenience, we focus on the ADR analysis by assuming $\hat{S}_1 = S_1$, following the spirit of model parsimony. If $S_1 \subset \hat{S}_1 \subset S_1 \cup S_2$, a similar analysis can be carried out by redefining $(S_1, S_2) = (\hat{S}_1, \hat{S}_1 \cap S_2)$, as discussed in Section 4.1. Together with the results in Theorem 1, such that $P(\hat{S}_3 = S_3) \to 1$, S_3 is also removed from the PSE with a large probability. Thus, the risk analysis in this section will be conducted by assuming both S_1 and S_3 are known in advance.

Definition 1. For any estimator $\boldsymbol{\beta}_{1n}^{\diamond}$ and $p_1 n -$ dimensional vector, $\mathbf{d}_1 n$, satisfying $\|\mathbf{d}_1 n\| \leq 1$, the ADR of $\mathbf{d}_{1n}' \boldsymbol{\beta}_{1n}^{\diamond}$ is

$$\mathrm{ADR}\left(\mathbf{d}_{1n}' \boldsymbol{\beta}_{1n}^{\diamond}\right) = \lim_{n \to \infty} E\left\{ \left[n^{1/2} s_{1n}^{-1} \mathbf{d}_{1n}' (\boldsymbol{\beta}_{1n}^{\diamond} - \boldsymbol{\beta}_1^*) \right]^2 \right\}, \quad (4.7)$$

where $s_{1n}^2 = \sigma^2 \mathbf{d}_{1n}' \boldsymbol{\Sigma}_{n11.2}^{-1} \mathbf{d}_{1n}$ with $\boldsymbol{\Sigma} n_{11.2}$ defined in (2.4).

We will provide some analytic expressions of ADRs under specific weak coefficients in (A2”). In particular,

(A2”) $\beta_j^* = \delta_j / \sqrt{n}$ for $j \in S_2$, where $|\delta_j| < \delta_{\max}$ for some $\delta_{\max} > 0$.

Denote $\boldsymbol{\delta} = \left(\delta_1, \ldots, \delta_{p_{2n}}\right)' \in \mathcal{R}^{p_{2n}}$. Then, $\Delta_n = \boldsymbol{\delta}' \boldsymbol{\Sigma}_{n22.1} \boldsymbol{\delta} \leqslant \rho_2 p_{2n} \delta_{\max}$, where ρ_2 is defined in (B4).

Define

$$\Delta_{\mathbf{d}_{1n}} = \frac{\mathbf{d}_{1n}'\left(\boldsymbol{\Sigma}_{n11}^{-1}\boldsymbol{\Sigma}_{n12}\boldsymbol{\delta}\boldsymbol{\delta}'\boldsymbol{\Sigma}_{n21}\boldsymbol{\Sigma}_{n11}^{-1}\right)\mathbf{d}_{1n}}{\mathbf{d}_{1n}'\left(\boldsymbol{\Sigma}_{n11}^{-1}\boldsymbol{\Sigma}_{n12}\boldsymbol{\Sigma}_{n22.1}^{-1}\boldsymbol{\Sigma}_{n21}\boldsymbol{\Sigma}_{n11}^{-1}\right)\mathbf{d}_{1n}}. \tag{4.8}$$

We obtain the following results on the expression of ADRs of PSEs.

Theorem 3. Let $\mathbf{d}_1 n$ be any $p_1 n -$ dimensional vector satisfying $0 < \|\mathbf{d}_1 n\| \leq 1$ and $s_{1n}^2 = \sigma^2 \mathbf{d}_{1n}' \boldsymbol{\Sigma}_{n11.2}^{-1} \mathbf{d}_{1n}$. Suppose all assumptions in Theorem 2 hold except that (A2) is replaced by (A2”). Then, we have

$$\mathrm{ADR}\left(\mathbf{d}_{1n}'\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}\right) = 1, \tag{4.9a}$$

$$\mathrm{ADR}\left(\mathbf{d}_{1n}'\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right) = 1 - (1-c)(1-\Delta_{\mathbf{d}_{1n}}), \tag{4.9b}$$

$$\mathrm{ADR}\left(\mathbf{d}_{1n}'\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{SE}}\right) = 1 - E[g_1(\mathbf{z}_2 + \boldsymbol{\delta})], \tag{4.9c}$$

$$\mathrm{ADR}\left(\mathbf{d}_{1n}'\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{PSE}}\right) = 1 - E[g_2(\mathbf{z}_2 + \boldsymbol{\delta})]. \tag{4.9d}$$

Here, $c = \lim_{n\to\infty} \mathbf{d}_{1n}'\boldsymbol{\Sigma}_{n11}^{-1}\mathbf{d}_{1n} / \left(\mathbf{d}_{1n}'\boldsymbol{\Sigma}_{n11.2}^{-1}\mathbf{d}_{1n}\right) \leqslant 1$, $\mathbf{z}_2$ satisfies that $s_{2n}^{-1}\mathbf{d}_{2n}'\mathbf{z}_2 \to N(0,1)$ with $\mathbf{d}_{2n} = \sigma^2\boldsymbol{\Sigma}_{n21}\boldsymbol{\Sigma}_{n11}^{-1}\mathbf{d}_{1n}$ and $s_{2n}^2 = \mathbf{d}_{2n}'\boldsymbol{\Sigma}_{n22.1}^{-1}\mathbf{d}_{2n}$. In addition,

$$g_1(\mathbf{x}) = \lim_{n\to\infty}(1-c)\frac{p_{2n}-2}{\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}}\left[2 - \frac{\mathbf{x}'((p_{2n}+2)\mathbf{d}_{2n}\mathbf{d}_{2n}')\mathbf{x}}{s_{2n}^2\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}}\right], \tag{4.10}$$

and

$$\begin{aligned} g_2(\mathbf{x}) = {} & \lim_{n\to\infty}\frac{p_{2n}-2}{\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}}\left[(1-c)\left(2 - \frac{\mathbf{x}'((p_{2n}+2)\mathbf{d}_{2n}\mathbf{d}_{2n}')\mathbf{x}}{s_{2n}^2\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}}\right)\right] I\left(\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x} \geqslant p_{2n}-2\right) \\ & + \lim_{n\to\infty}\left[(2 - s_{2n}^{-2}\mathbf{x}'\mathbf{d}_{2n}\mathbf{d}_{2n}'\mathbf{x})(1-c)\right] I\left(\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x} \leqslant p_{2n}-2\right), \end{aligned} \tag{4.11}$$

with $I(\cdot)$ being an indicator function.

Theorem 3 lists the analytic expressions of the asymptotic risk of all above estimators. From Theorem 3, we can obtain the following risk comparisons.

Corollary 3. Under assumptions in Theorem 3, we have

i. $\mathrm{ADR}\left(\mathbf{d}_{1n}'\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{PSE}}\right) \leqslant \mathrm{ADR}\left(\mathbf{d}_{1n}'\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{SE}}\right) \leqslant \mathrm{ADR}\left(\mathbf{d}_{1n}'\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}\right)$ holds for $0 < \|\boldsymbol{\delta}\|^2 \leq 1$;

ii. Inequalities in (i) also hold for $\|\boldsymbol{\delta}\|^2 \leq 1 + \iota$ for some $\iota > 0$ if $\Delta n = \iota p_2 n$.

iii. If $\|\boldsymbol{\delta}\| = o(1)$, then $\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n}\right) \leqslant \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{PSE}}_{1n}\right) < \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{1n}\right)\}$ holds for $\boldsymbol{\delta} = 0$, where the '=' holds when $p_2 n \to \infty$.

Corollary 3 indicates that the performance of the PSE is closely related to the post selection least squares estimator. On one hand, if $\hat{S}_1 \subset S_1 \cup S_2$ and $(S_1 \cup S_2) \cap \hat{S}_1^c$ are large, then the post selection PSE tends to dominate the RE. Thus, the post selection PSE can improve the performance of the post selection least squares estimators in [13] and [14], especially when $pn \gg n$ and an under-fitted submodel is selected by a large penalty parameter. On the other hand, if a variable selection approach almost generates the right submodel and $\|\boldsymbol{\delta}\| = o(1)$, that is, $\lim_{n\to\infty} \hat{S}_1 = S_1 \cup S_2$, then a post selection LSE ($\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n}$) is the most efficient one compared with all other post selection estimates.

Remark 6. In the high-dimensional setting where $p \gg n$, we do need to assume the true model to be sparse in the sense that most coefficients goes to 0 when $n \to \infty$. However, we still permit some βj to be small but not exactly 0. Such covariates with a small amount of influence on the response variable are often ignored incorrectly in high-dimensional variable selection methods. If we borrow information from those covariates using the proposed shrinkage methods, the prediction performance based on selected submodel can be improved substantially.

5 Simulation studies

In this section, we use some simulation studies to examine the quadratic risk performance of the proposed estimators. Our simulation is based on the linear regression model in (1.1).

True model setting. In all experiments, ϵi's are simulated from independent and identically distributed standard normal random variables, $x_{is} = \left(\xi^1_{(is)}\right)^2 + \xi^2_{(is)}$, where $\xi^1_{(is)}$ and $\xi^2_{(is)}, i = 1, \cdots, n$, $s = 1, \cdots, pn$ are also independent copies of the standard normal distribution. In all experiments, we let $n = 200$ and $pn = n\tau$ for different sample size n, where τ changes from 1 to 1.2 with an increment of 0.02. Three different coefficient configurations are considered as follows:

Case 1: $\boldsymbol{\beta}^* = (5, 5, 5, \underbrace{0.5, \ldots, 0.5}_{10}, \mathbf{0}'_{p_3})'$;

Case 2: $\boldsymbol{\beta}^* = (10, 10, 10, \underbrace{0.1, \ldots, 0.1}_{50})', \mathbf{0}'_{p_3})'$;

Case 3: $\boldsymbol{\beta}^* = (10, 10, 10, \underbrace{0.1, \ldots, 0.1}_{p_2}, \mathbf{0}'_{20})'$.

All nonzero coefficients are randomly assigned to be either positive or negative. Both zero and weak signals coexist in the aforementioned three settings. In Case 1, most covariates are noises. Compared with Case 1, the weak signals become weaker, and the strong signals become stronger

in Case 2. In addition, the number of weak signals is larger but also fixed. In Case 3, only $p_3n = 20$ zero signals, large amount of weak signals contribute simultaneously, and the number of weak signals grows with the number of covariates such that $p_2n \gg n$. Notice that the signal strength setting in this case is different from that considered in our post selection shrinkage analysis, where $p_2n < n$ and $p_3n \gg n$.

Subset selection. Because the adaptive Lasso, smoothly clipped absolute deviation, and minimax concave penalty perform closely under certain conditions, we only adopt the adaptive Lasso, and Lasso in selecting a subset before applying the post selection shrinkage strategy. All tuning parameters in variable selection approaches are chosen using the BIC.

Tuning parameters and simulation Setting. As we know, *an* and *rn* are two important tuning parameters affecting $\hat{S}_2$ and $\hat{S}_3$. We choose those two tuning parameters based upon the asymptotic investigations in Theorem 2 for all our numerical studies. In particular, the post selection PSEs are computed for $r_n = c_2 a_n^{-2} (\log\log n)^3 \log(n \vee p_n)$ with $an = c_1 n^{-1/8}$. Corresponding coefficients c_1 and c_2 are chosen using cross validation.

Evaluation. Each design is repeated 1000 times, as a further increase in the number of realizations did not significantly change the result. Let $\boldsymbol{\beta}_{1n}^{\diamond}$ be either $\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{PSE}}$ or $\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}$ after the variable selection. The performance of $\boldsymbol{\beta}_{1n}^{\diamond}$ is evaluated by the relative mean squared error (RMSE) criterion with respect to $\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}$ as follows:

$$\mathrm{RMSE}(\boldsymbol{\beta}_{1n}^{\diamond}) = \frac{E\|\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^*\|^2}{E\|\boldsymbol{\beta}_{1n}^{\diamond} - \boldsymbol{\beta}_1^*\|^2}. \quad (5.1)$$

Therefore, $\mathrm{RMSE}(\boldsymbol{\beta}_{1n}^{\diamond}) > 1$ means the superiority of $\boldsymbol{\beta}_{1n}^{\diamond}$ over $\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}$.

Result: We plot the mean RMSEs from 1000 iterations along *pn* in Figure 1. Some selected results are also reported in Table 1. To check the behavior of Lasso or adaptive Lasso for subset selection, we also report the average number of selected important covariates as $|\hat{S}_1|$ in Table 1. It is not surprising to see that RE post the adaptive Lasso is comparable with the adaptive Lasso itself, while RE post the Lasso behaves much better than Lasso [13, 14]. We summarize the simulation results as follows:

- Figure 1(a')–(c') lists results when the adaptive Lasso is used to generate the submodel. (i) When *pn* is closer to *n*, both post selection RE and adaptive Lasso perform better than the post selection PSE and WR (RMSE > 1). (ii) When *pn* grows bigger, both RE and adaptive Lasso become worse than the post selection WR (RMSE < 1). However, the post selection PSE still performs better than the post selection WR. Therefore, the post selection PSE provides a protection of the adaptive Lasso in the case that the adaptive Lasso loses its efficiency.

- Figure 1(a)–(c) lists results when the Lasso is used to generate the submodel. The advantage of the post selection PSE over the Lasso is more obvious than the earlier. This is because the adaptive Lasso tends to produce a more efficient estimator than the Lasso does.
- When *pn* grows, the post selection PSE is much more robust and at least as good as the WR estimator (RMSE is approaching to 1). When *pn* grows bigger, the improvement of the post selection PSE from adaptive Lasso or Lasso becomes more obvious. See Table 1.
- In Case 3, the post selection PSE may lose its superiority to the post selection RE and adaptive Lasso, especially when *pn* grows quickly with *n*. One explanation is that the selected model size varies dramatically because the number of weak coefficients grows. However, if we still follow the model parsimony spirit and decide to use an aggressive tuning parameter to obtain a relatively consistent submodel size $\hat{S}_1$, the superiority of post selection PSEs follows the same pattern as in Cases 1 and 2.

Table 1. Simulated RMSEs from simulation examples in Case 1–3.

		Lasso					**Adaptive Lasso**			
Case	***pn***	$\lvert\hat{S}_1\rvert$	$\hat{\beta}_{1n}^{\text{Lasso}}$	$\hat{\beta}_{1n}^{\text{RE}}$	$\hat{\beta}_{1n}^{\text{PSE}}$		$\lvert\hat{S}_1\rvert$	$\hat{\beta}_{1n}^{\text{ALasso}}$	$\hat{\beta}_{1n}^{\text{RE}}$	$\hat{\beta}_{1n}^{\text{PSE}}$
	200	10.920	0.690	7.880	2.285		10.537	7.611	7.739	2.269
	222	10.785	0.190	2.035	1.680		10.434	2.001	1.991	1.667
1	275	10.655	0.082	0.744	1.231		10.250	0.783	0.773	1.242
	340	10.491	0.066	0.574	1.126		10.137	0.585	0.558	1.114
	420	10.416	0.061	0.485	1.061		9.906	0.514	0.491	1.062
	519	10.293	0.063	0.476	1.047		9.781	0.480	0.446	1.042
	200	3.112	0.491	6.169	2.409		3.170	4.859	3.431	2.199
	222	3.078	0.137	1.790	1.807		3.149	1.447	1.012	1.640
2	275	3.041	0.048	0.684	1.393		3.083	0.561	0.384	1.205
	340	3.036	0.035	0.517	1.222		3.051	0.395	0.270	1.066
	420	3.000	0.029	0.442	1.138		3.025	0.335	0.233	1.003
	519	3.000	0.023	0.388	1.140		3.000	0.312	0.217	0.998
	200	4.020	0.730	2.594	1.420		7.379	6.380	5.815	1.491
	222	6.109	0.430	0.809	1.200		10.005	1.778	1.684	1.310
3	275	5.277	0.176	0.449	1.007		8.159	0.747	0.687	1.092
	340	3.046	0.034	0.396	1.077		3.783	0.476	0.361	1.070
	420	5.325	0.231	0.633	0.984		7.390	0.762	0.710	1.025
	519	7.213	0.461	0.860	1.014		9.114	0.844	0.804	1.020

|Ŝ1| is the average size of produced submodel; RMSEs, relative mean squared errors;PSE, post selection shrinkage estimator; RE, restricted estimator.

Figure 1. Relative mean squared errors (RMSEs) of post selection relative mean squared errora (PSEs) compared with one from Lasso or adaptive Lasso (ALasso) from simulation examples in Cases 1–3. The top (a or a'), middle (b or b'), and bottom (c or c') panels are for Cases 1, 2, and

3, respectively. The left (a–c) and right panels (a'–c') are comparisons when the candidate submodels are chosen from the Lasso and adaptive Lasso methods, respectively.

6 Real-data example

In this section, we apply the proposed post selection shrinkage strategy to the growth data for the years 1960–1985 [30]. Table 2 lists the detailed descriptions of the dependent variable and 45 covariates related to education and its interaction with lgdp60*i*, market efficiency, political stability, market openness, and demographic characteristics.

Table 2. List of variable.

Variable	Description
Dependent variable	
gr	Annualized GDP growth rate in the period of 1960–85
Threshold variables	
gdp60	Real GDP per capita in 1960 (1985 price)
Covariates	
lgdp60	log GDP per capita in 1960 (1985 price)
lsk	Log(Investment/Output) annualized over 1960–85;
	a proxy for the log physical savings rate
lgrpop	Log population growth rate annualized over 1960–1985
pyrm60	Log average years of primary schooling in the male population in 1960
pyrf60	Log average years of primary schooling in the female population in 1960
syrm60	Log average years of secondary schooling in the male population in 1960
syrf60	Log average years of secondary schooling in the female population in 1960
hyrm60	Log average years of higher schooling in the male population in 1960
hyrf60	Log average years of higher schooling in the female population in 1960
nom60	Percentage of no schooling in the male population in 1960
nof60	Percentage of no schooling in the female population in 1960
prim60	Percentage of primary schooling attained in the male population in 1960
prif60	Percentage of primary schooling attained in the female population in 1960
pricm60	Percentage of primary schooling complete in the male population in 1960
pricf60	Percentage of primary schooling complete in the female population in 1960
secm60	Percentage of secondary schooling attained in the male population in 1960
secf60	Percentage of secondary schooling attained in the female population in 1960
seccm60	Percentage of secondary schooling complete in the male population in 1960
seccf60	Percentage of secondary schooling complete in the female population in 1960
llife	Log of life expectancy at age 0 averaged over 1960–1985
lfert	Log of fertility rate (children per woman) averaged over 1960–1985
edu/gdp	Government expenditure on eduction per GDP averaged over 1960–1985
gcon/gdp	Government consumption expenditure net of defence and education
	per GDP averaged over 1960–85
revol	The number of revolutions per year over 1960–84
revcoup	The number of revolutions and coups per year over 1960–84

wardum	Dummy for countries that participated in at least one external war over 1960–84
wartime	The fraction of time over 1960–1985 involved in external war
lbmp	Log(1+black market premium averaged over 1960–85)
tot	The term of trade shock
lgdp60	'educ' Product of two covariates (interaction of lgdp60 and education
	variables from pyrm60 to seccf60); total 16 variables

The growth regression model has been applied to test the negative relationship between the long-run growth rate and the initial GDP given other covariates. See [31] and [32] for literature reviews. Very recently, [33] took into account the possible discrepancy among the aforementioned negative relationship using a growth regression model with threshold. In particular, they consider a threshold variable in the following regression model,

$$\mathrm{gr}_i = \beta_0 + \beta_1 lgdp60_i + \mathbf{z}_i'\boldsymbol{\beta}_2 + I(Q_i < \tau)(\delta_0 + \delta_1 lgdp60_i + \mathbf{z}_i'\boldsymbol{\delta}_2) + \varepsilon_i, \quad (6.1)$$

where gr*i* is the annualized GDP growth rate of country *i* from 1960 to 1985, lgdp60*i* is the log GDP in 1960, **z***i* includes all 45 covariates listed in Table 2, and *Qi* is a threshold variable, where we use the initial GDP in 1960. Because the estimation of the threshold parameter τ is not our target, we consider five different τ's in our analysis: 1655, 2073, 2898, 3268, and 6030. Among them, $\tau = 2898$ is a threshold value suggested by [33], and the other four threshold values are *k*th percentiles for $k = 60,70,80,90$, respectively. After removing all missing data, each setting includes $n = 82$ observations and $p = 90$ covariates besides two intercepts.

Before applying the post selection shrinkage strategy, we first obtain candidate subsets from two variable selection techniques: Lasso and adaptive Lasso, respectively. All tuning parameters are selected from fivefold cross validation. In Table 3, we list the numbers of selected important variables, $\hat{s}_1 = |\hat{S}_1|$, and also the sizes of candidate submodels, under five different τ's. In Table 4, we list the frequency of each variable being selected among all five settings. We observe that Lasso and adaptive Lasso variable selection results in Table IV are quite close for this data set. However, the selected candidate subset model can be quite different among all five different τ's.

Table 3. Sizes of selected submodel.

τ	6030	3268	2898	2073	1655
Lasso	15	18	18	19	11
adaptive Lasso	19	13	20	19	11

Table 4. Frequency of selected variables (based upon either $\beta j \neq 0$ or $\delta j \neq 0$) among All 5 τ's.

Variable	Lasso		ALasso	
	#($\beta j \neq 0$)	#($\delta j \neq 0$)	#($\beta j \neq 0$)	#($\delta j \neq 0$)
lgdp60	5	0	5	0
lsk	5	0	5	0

nom60	0	1	0	1
prim60	3	0	3	0
pricm60	3	3	3	3
seccm60	0	5	0	5
seccf60	1	0	1	0
llife	5	0	5	0
lfert	5	0	5	0
edugdp	3	0	4	0
gcongdp	5	0	5	0
revol	2	0	3	0
wardum	2	3	2	3
wartime	4	4	3	3
lbmp	5	0	5	0
tot	0	5	0	5
lgdpsyrm60	2	0	2	0
lgdphyrm60	3	0	1	0
lgdphyrf60	0	1	1	0
lgdpnof60	0	3	0	3
lgdpprim60	2	0	2	1
lgdpprif60	0	1	0	2
lgdpseccf60	1	0	0	0

After the variable selection, post selection PSE is applied based upon the selected candidate subsets in all settings. Tables 5 and 6 give the estimation results for $\tau = 2898$ and $\tau = 1655$, where both candidate subsets are selected by the adaptive Lasso. We omit results under other settings to save the space.

Table 5. Estimation results under $\tau = 2898$ (Candidate submodel from ALasso).

Variable	$\hat{\beta}^{(ALasso)}$	$\hat{\delta}^{(ALasso)}$	$\hat{\beta}^{(PSE)}$	$\hat{\delta}^{(PSE)}$
lgdp60	-9.253×10^{-3}	—	-1.287×10^{-2}	—
lsk	6.121×10^{-4}	—	3.942×10^{-4}	—
nom60	—	1.400×10^{-2}	—	3.481×10^{-2}
prim60	-4.579×10^{-2}	—	-7.472×10^{-2}	—
pricm60	1.934×10^{-2}	1.974×10^{-3}	4.129×10^{-2}	7.058×10^{-3}
seccm60	—	4.903×10^{-4}	—	4.324×10^{-4}
llife	1.200×10^{-3}	—	2.212×10^{-3}	—
lfert	-1.659×10^{-3}	—	-1.507×10^{-3}	—
edugdp	2.228×10^{-5}	—	2.309×10^{-5}	—
gcongdp	-2.351×10^{-4}	—	-2.610×10^{-4}	—
revol	-1.020×10^{-6}	—	-1.158×10^{-4}	—
wardum	—	-1.417×10^{-4}	—	-3.336×10^{-4}
wartime	-1.655×10^{-4}	—	-5.081×10^{-5}	—
lbmp	-1.580×10^{-3}	—	-1.595×10^{-3}	—
tot	—	5.202×10^{-6}	—	6.318×10^{-6}
lgdphyrm60	1.122×10^{-2}	—	4.291×10^{-2}	—

lgdphyrf60	-7.585×10^{-3}	—	-4.143×10^{-2}	—
lgdpnof60	—	6.392×10^{-2}	—	0.189
lgdpprif60	—	-3.130×10^{-2}	—	−0.127

Table 6. Estimation results under τ = 1655 (Candidate submodel from ALasso).

Variable	$\hat{\beta}^{(ALasso)}$	$\hat{\delta}^{(ALasso)}$	$\hat{\beta}^{(PSE)}$	$\hat{\delta}^{(PSE)}$
lgdp60	-2.841×10^{-3}	—	-1.306×10^{-2}	—
lsk	1.319×10^{-3}	—	1.284×10^{-3}	—
seccm60	—	3.652×10^{-4}	—	5.873×10^{-4}
llife	3.532×10^{-4}	—	1.633×10^{-3}	—
lfert	-2.552×10^{-4}	—	-2.250×10^{-3}	—
gcongdp	-1.554×10^{-4}	—	-3.033×10^{-4}	—
revol	-3.715×10^{-5}	—	-9.248×10^{-4}	—
wartime	-4.965×10^{-5}	-1.120×10^{-5}	2.731×10^{-4}	-3.958×10^{-5}
lbmp	-1.428×10^{-3}	—	-5.887×10^{-4}	—
tot	—	5.175×10^{-7}	—	8.476×10^{-6}

Becuase we do not know what the true model is in the real-data analysis, we first evaluate the prediction improvement from variable selection estimates to post selection PSEs by computing the relative residual sum of squares (RRSS) of the estimator $\boldsymbol{\beta}_{\mathcal{J}}^{\diamond}$ over the WR estimator $\hat{\boldsymbol{\beta}}_{\mathcal{J}}^{\mathrm{WR}}$ as follows:

$$\mathrm{RRSS}(\boldsymbol{\beta}_{\mathcal{J}}^{\diamond}) = \frac{\sum_{i=1}^{n} \|\mathbf{y} - \sum_{j\in\mathcal{J}} \mathbf{X}_{\mathcal{J}} \hat{\boldsymbol{\beta}}_{\mathcal{J}}^{\mathrm{WR}}\|^2}{\sum_{i=1}^{n} \|\mathbf{y} - \sum_{j\in\mathcal{J}} \mathbf{X}_{\mathcal{J}} \boldsymbol{\beta}_{\mathcal{J}}^{\diamond}\|^2}, \quad (6.2)$$

where $\mathcal{J}$ is the index of the submodel chosen by corresponding variable selection methods, and $\boldsymbol{\beta}_{\mathcal{J}}^{\diamond}$ can be (adaptive) Lasso and the corresponding generated post selection SEs and post selection PSEs. Similar to the simulation studies, RRSS > 1 indicates the superiority of $\boldsymbol{\beta}_{\mathcal{J}}^{\diamond}$ over $\hat{\boldsymbol{\beta}}_{\mathcal{J}}$. The results on RRSS for different τ's are reported in Figure 2, where the left and right panels are based upon Lasso and adaptive Lasso submodels, respectively. Those RRSS values of post selection REs give the highest value in both cases. This is not surprising because we assume the selected submodel is the right one and does not account for any bias. In both cases, the post selection PSEs dominate the corresponding variable selection estimation in terms of the RRSS regardless of whether Lasso or adaptive Lasso is used for generating the candidate submodel. This is because shrinkage estimation provides a better trade-off between bias and variance when selected submodels underfit the true model.

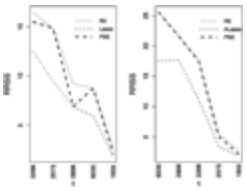

Figure 2. Relative residual sum of squares (RRSS) from (6.2) from post selection post selection shrinkage estimator (PSE) and the Lasso-type estimators: Lasso (left panel) or adaptive Lasso (ALasso) (right panel). The curve is plotted based upon a decreasing order of RRSS for better visibility, with corresponding values of τ plotted in x-axis.

In addition, we also obtain prediction errors using cross validation following 500 random partitions of the data set. In each partition, the training set consists of 2/3 observations (size 55), and the test set consists of the remaining 1/3 observations (size 28). Corresponding results for $\tau = 2898$ and 1655 are reported in Figure 3, where the post selection PSEs are compared with the adaptive Lasso. The comparisons between the post selection PSEs and (adaptive) Lasso for other τ's follow the similar pattern and thus are omitted. It is observed that post selection PSEs produce much smaller prediction errors than the Lasso-type estimation.

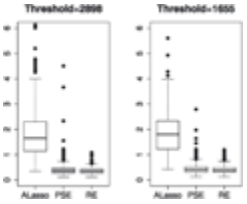

Figure 3. Prediction errors from post selection post selection shrinkage estimator (PSE), restricted estimator (RE), and adaptive Lasso (ALasso). Left: $\tau = 2898$; Right: $\tau = 1655$. All prediction errors are computed using cross validation following 500 random partitions of the data set. In each partition, the training set consists of 2/3 observations, and the test set consists of the remaining 1/3 observations.

7 Conclusion and discussions

In this paper, we generalize the shrinkage estimation to a high-dimensional sparse regression model. We propose a post selection shrinkage estimation strategy by shrinking a WR estimator in the direction of a candidate submodel obtained by existing PLSs variable selection methods.

When pn grows with n quickly, it is reasonable to assume that the model sparsity exists in the sense that most covariates do not contribute. However, at the same time, some covariates may still make some small but jointly non-trivial contribution to the response. Existing penalized regularization approaches usually lead to a sparse model but tend to miss the possible small contributions from some covariates, resulting in excessive prediction errors or inefficient estimation. Our proposed post selection shrinkage strategy, taking into account possible contributions of covariates with weak and/or moderate signals, has dominant prediction performances over candidate submodel estimates generated from Lasso-type methods.

Before obtaining a shrinkage estimator, one key step is to generate a full estimation of $\boldsymbol{\beta}n$ when $p \gg n$. We suggest a post selection WR estimator which is able to separate small coefficients from zero coefficients. The advantages of proposed post selection PSE are studied both theoretically and numerically. In theory, we established the asymptotic normality of the post selection WR estimator when pn grows with n at an almost exponential rate such that $\log(pn) = O(nv)$ for some $0 < v < 1$. Those novel asymptotic properties are used for investigating the asymptotic efficiency of the proposed post selection PSE analytically. In numerical studies, we chose tuning parameters c_1 and c_2 from cross validation but cannot guarantee their optimality for post selection PSE. The choice of tuning parameters is an important but challenging issue in high-dimensional data analysis that could potentially create very important future work. Although the proposed post selection PSE was presented based on a WR method, other methods can also be used to generate the shrinkage estimator.

Finally, we acknowledge the importance of Lasso-type variable selection methods, but at the same time, and do not depend completely on them, especially when many weak coefficients jointly affect the response variable. The Lasso is the start but not the end. We could potentially still make some significant prediction improvements. We hope this work will shed some more light on the investigation of the post variable selection shrinkage analysis in high-dimensional data analysis.

Appendix

All technical proofs are given in this section.

Proof of Theorem 1. After solving (3.1), we obtain

$$\tilde{\boldsymbol{\beta}}_{\hat{S}_1}(r_n) = \left(\mathbf{X}_{\hat{S}_1}{}'\boldsymbol{M}_{\hat{S}_1^c}(r_n)\mathbf{X}_{\hat{S}_1}\right)^{-1}\mathbf{X}'_{\hat{S}_1}\boldsymbol{M}_{\hat{S}_1^c}(r_n)\mathbf{y} \quad \text{(A1)}$$

and

$$\tilde{\boldsymbol{\beta}}_{\hat{S}_1^c}(r_n) = \left(r_n\mathbf{I}_{p_{2n}} + \mathbf{X}_{\hat{S}_1^c}{}'\boldsymbol{M}_{\hat{S}_1}\mathbf{X}_{\hat{S}_1^c}\right)^{-1}\mathbf{X}'_{\hat{S}_1^c}\boldsymbol{M}_{\hat{S}_1}\mathbf{y}, \quad \text{(A2)}$$

where $\boldsymbol{M}_{\hat{S}_1^c}(r_n) = \mathbf{I}_n - \mathbf{X}_{\hat{S}_1^c}\left(r_n\mathbf{I}_{p_{2n}} + \mathbf{X}'_{\hat{S}_1^c}\mathbf{X}_{\hat{S}_1^c}\right)^{-1}\mathbf{X}'_{\hat{S}_1^c}$ and $\boldsymbol{M}_{\hat{S}_1} = \mathbf{I}_n - \mathbf{X}_{\hat{S}_1}(\mathbf{X}'_{\hat{S}_1}\mathbf{X}_{\hat{S}_1})^{-1}\mathbf{X}'_{\hat{S}_1}$.

We only need to prove the result under the condition $\hat{S}_1 = S_1$, and then all matrices, vectors indexed by $\hat{S}_1$ can be replaced by S_1 or 1 without causing of any confusion. For example, $\boldsymbol{M}_{\hat{S}_1} = \boldsymbol{M}_{S_1} = \boldsymbol{M}_1$ under the condition.

First, we check the bias of $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{S_1^c}$. Because $\boldsymbol{M}_1$ is an idempotent matrix, $\boldsymbol{M}_1\mathbf{X}_1 n = 0$. Denote $qn = p_2 n + p_3 n$. Then,

$$(\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n})^{-1}\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{1n}\boldsymbol{\beta}_{10} = \mathbf{0}.$$

Let $\mathbf{Q}$ be a $qn \times qn$ orthogonal matrix such that

$$\mathbf{U}'\boldsymbol{M}_1\mathbf{U} = \mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c} = \mathbf{Q}\begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\mathbf{Q}',$$

where $\mathbf{D} = \operatorname{diag}\{\varrho_{1n}, \ldots, \varrho_{k_n n}\}$. Then, we have

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{S_1^c}\right) - \boldsymbol{\beta}^*_{S_1^c} &= \left(\mathbf{X}_{S_1^c}{}'\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n}\right)^{-1}\mathbf{X}_{S_1^c}{}'\boldsymbol{M}_1\mathbf{y} - \boldsymbol{\beta}^*_{S_1^c}\Big] \\ &= \left(\mathbf{X}_{S_1^c}{}'\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n}\right)^{-1}\mathbf{X}_{S_1^c}{}'\boldsymbol{M}_1\mathbf{X}_{S_1^c}\boldsymbol{\beta}^*_{S_1^c} - \boldsymbol{\beta}^*_{S_1^c} \\ &= -r_n\left(\mathbf{X}_{S_1^c}{}'\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n}\right)^{-1}\boldsymbol{\beta}^*_{S_1^c} \\ &= -\mathbf{Q}\begin{pmatrix} \left(\mathbf{I}_{k_n} + r_n^{-1}\mathbf{D}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{q_n - k_n} \end{pmatrix}\mathbf{Q}'\boldsymbol{\beta}^*_{S_1^c}. \end{aligned} \tag{A3}$$

Suppose that $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ and $\mathbf{Q}_1$ is a $qn \times kn$ matrix. Notice that $\mathbf{Q}\mathbf{Q}' = \mathbf{Q}'\mathbf{Q} = \mathbf{I}_{q_n}$. Then, $\mathbf{Q}_1'\mathbf{Q}_1 = \mathbf{I}_{k_n}$, $\mathbf{Q}1'\mathbf{Q}_2 = \mathbf{0}$, and $\mathbf{Q}_2\mathbf{Q}2'$ is a projection matrix. Let $\boldsymbol{\theta}^* = \mathbf{Q}_1\mathbf{Q}_1'\boldsymbol{\beta}^*_{S_1^c}$. Then,

$$\boldsymbol{\beta}^*_{S_1^c} = \mathbf{Q}_1\mathbf{Q}_1{}'\boldsymbol{\theta}^* = \mathbf{Q}_1\mathbf{Q}_1{}'\mathbf{Q}_1\mathbf{Q}_1{}'\boldsymbol{\beta}^*_{S_1^c} = \mathbf{Q}_1\mathbf{Q}_1{}'\boldsymbol{\beta}^*_{S_1^c}. \tag{A4}$$

Replace $\boldsymbol{\beta}^*_{S_1^c}$ in (A3) by $\mathbf{Q}_1\mathbf{Q}_1'\boldsymbol{\beta}^*_{S_1^c}$, we have

$$E\left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{S_1^c}\right) - \boldsymbol{\beta}_{0S_1^c} = -\mathbf{Q}_1\left(\mathbf{I}_{k_n} + r_n^{-1}\mathbf{D}\right)^{-1}\mathbf{Q}_1'\boldsymbol{\beta}^*_{S_1^c}.$$

Thus,

$$\left\|E\left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{S_1^c}\right) - \boldsymbol{\beta}*_{S_1^c}\right\|^2 = \boldsymbol{\theta}_0'\mathbf{Q}_1\left(\mathbf{I}_{k_n} + r_n^{-1}\mathbf{D}\right)^{-2}\mathbf{Q}_1\boldsymbol{\theta}_0 \leqslant (1 + \varrho_{1n}/r_n)^{-2}\|\boldsymbol{\beta}_{0S_1^c}\|^2$$

For every $j \notin S_1$, $\left|\operatorname{bias}\left(\hat{\beta}_j^{\mathrm{WR}}\right)\right| \leqslant \left\|E\left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{S_1^c}\right) - \boldsymbol{\beta}_{0S_1^c}\right)\right\|$, and thus,

$$\left|\operatorname{bias}\left(\hat{\beta}_j^{\mathrm{WR}}\right)\right| \leqslant (1 + \varrho_{1n}/r_n)^{-1}\|\boldsymbol{\beta}_{0S_1^c}\| \leqslant (r_n/\varrho_{1n})O(n^{\tau}) \leqslant O(r_n n^{\tau-\eta}).$$

The rest of the proof just mimics the proof of Theorem 2 in [28]. We will provide some outlines of the proof. If we let $r_n = c_2 a_n^{-2}(\log\log n)^3 \log(n \vee p)$ and $\log(pn) = O(nv)$ in (B3), then for $un = 1 + (\log\log n)^{-1}$, we have

$$\frac{\left|\operatorname{bias}\left(\hat{\beta}_j^{\mathrm{WR}}\right)\right|}{a_n(u_n - 1)} \leqslant \frac{r_n n^{\tau-\eta}}{a_n(u_n - 1)} \leqslant \frac{c_2(\log\log n)^4}{a_n^3 n^{\eta-\tau-v}} \leqslant \frac{c_2(\log\log n)^4}{c_1^3 n^{\eta-\tau-v-3\alpha}} \to 0 \quad \text{if } 3\alpha < \eta - v - \tau,$$

where the last '≤' is from (3.5) and c_1 is defined there. From the normal assumption of ϵi and the solution in linear expression in (A2), we know $\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{S_1^c}$ is normally distributed and

$$\begin{aligned}\mathrm{Var}\left(\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{S_1^c}\right) &= \sigma^2\left(\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n}\right)^{-1}\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c}\left(\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n}\right)^{-1}\\ &\preceq \sigma^2\left(\mathbf{X}'_{S_1^c}\boldsymbol{M}_1\mathbf{X}_{S_1^c} + r_n\mathbf{I}_{q_n}\right)^{-1}\\ &\preceq \sigma^2 r_n^{-1}\mathbf{I}_{q_n},\end{aligned}$$

where '$\mathbf{A}\preccurlyeq\mathbf{B}$' means $\mathbf{B}-\mathbf{A}$ is a non-negative definite matrix. Thus, for any $j \notin S_1$, $\mathrm{Var}\left(\hat{\beta}_j^{\mathrm{WR}}\right) = O(1/r_n)$. Notice that $\sqrt{r_n}a_n(u_n-1) = O\left((\log\log n)^{1/2}\right) \to \infty$. We have

$$\frac{a_n(u_n-1)}{\sqrt{\mathrm{Var}\left(\hat{\beta}_j^{\mathrm{WR}}\right)}} \geqslant a_n(u_n-1)\sqrt{r_n} \to \infty.$$

$$\begin{aligned}P\left(|\hat{\beta}_j^{\mathrm{WR}} - \beta_j^*| > a_n(u_n-1)\right) &\leqslant P\left(|N(0,1)| > \frac{a_n(u_n-1)}{\sqrt{\mathrm{Var}(\hat{\beta}_j^{\mathrm{WR}})}} - \frac{|\mathrm{bias}(\hat{\beta}_j^{\mathrm{WR}})|}{\sqrt{\mathrm{Var}(\hat{\beta}_j^{\mathrm{WR}})}}\right)\\ &= 2\Phi\left(\frac{|\mathrm{bias}(\hat{\beta}_j^{\mathrm{WR}})| - a_n(u_n-1)}{\sqrt{\mathrm{Var}(\hat{\beta}_j^{\mathrm{WR}})}}\right)\\ &\leqslant 2\Phi\left(-c_0\sqrt{r_n}a_n/(\log\log n)\right)\\ &\leqslant \exp\{-c_0^2 r_n a_n^2/(\log\log n)^2\},\end{aligned}$$

where Φ is the cumulative distribution function of a standard normal random variable, $c_0 > 0$ is a constant, '$\leq$' is the tail probability of a normal random variable. Thus,

$$
\begin{aligned}
&P\left(\{j \notin S_1 : |\beta_j^*| > a_n u_n\} \subset \{j \notin S_1 : |\hat{\beta}_j^{\mathrm{WR}}| > a_n\}\right) \\
&\geqslant 1 - P\left(\bigcup_{j:|\beta_j^*|>a_n u_n} \{|\hat{\beta}_j^{\mathrm{WR}}| \leqslant a_n\}\right) \\
&\geqslant 1 - P\left(\bigcup_{j:|\beta_j^*|>a_n u_n} \{|\hat{\beta}_j^{\mathrm{WR}} - \beta_j^*| \leqslant a_n(u_n - 1)\}\right) \\
&\geqslant 1 - \sum_{j \notin S_1} P\left(|\hat{\beta}_j^{\mathrm{WR}} - \beta_j^*| > a_n(u_n - 1)\right) \\
&\geqslant 1 - q_n \exp\{-c_0^2 r_n a_n^2/(\log\log n)^2\} \\
&\geqslant 1 - \exp\{-\left(c_0^2 r_n a_n^2/(\log\log n)^2 - \log(p_n)\right)\} \\
&\geqslant 1 - \exp\{-(c_0^2 \log\log n - 1)\log(p_n \vee n)\}.
\end{aligned}
$$

When n is large enough, there exists $c_0^2 \log\log n - 1 > t > 0$ for some $t > 0$. Thus,

$$
\lim_{n\to\infty} P\left(\{j \notin S_1 : |\beta_j^*| > a_n u_n\} \subset \{j \notin S_1 : |\hat{\beta}_j^{\mathrm{WR}}| > a_n\}\right) \geqslant 1 - (p_n \vee n)^{-t} \to 1.
$$

Similarly, we have

$$
\lim_{n\to\infty} P\left(\{j \notin S_1 : |\beta_j^*| > a_n/u_n\} \supset \{j \notin S_1 : |\hat{\beta}_j^{\mathrm{WR}}| > a_n\}\right) \geqslant 1 - (p_n \vee n)^{-t} \to 1.
$$

Because of the continuity of $\hat{\beta}_j^{\mathrm{WR}}$ and $\lim_{n\to\infty} u_n = 1$, we have

$$
\lim_{n\to\infty} P(\hat{S}_2 | \hat{S}_1 = S_1) = 1.
$$

Proof of Corollary 1. Because $S_1 \subset \hat{S}_1$, a weighted ridge estimator $\hat{\boldsymbol{\beta}}_{\hat{S}_1}$ aims to find some weak signals from $\hat{S}_1^c \cap S_2$. Because $\hat{S}_1^c \subset S_1^c$, the smallest positive eigenvalues of $\mathbf{X}'_{\hat{S}_1^c} \boldsymbol{M}_{\hat{S}_1} \mathbf{X}_{\hat{S}_1}$ must be larger than $\lambda_1 n$, and $\|\boldsymbol{\beta}^*_{\hat{S}_1^c}\|_2 \leqslant \|\boldsymbol{\beta}^*_{S_1^c}\|_2$. Thus, we can borrow the proof of Theorem 1 here, by treating $\hat{S}_1$ and $S_2 \cap \hat{S}_1^c$ as the new S_1 and S_2.

Proof of Theorem 2. Similar to the proof in Theorem 1, we assume $\hat{S}_1 = S_1$. Then, the penalized quadratic loss function in (3.1) becomes

$$
L(\boldsymbol{\beta}_n; S_1) = \left\{\|\mathbf{y} - \mathbf{X}_n \boldsymbol{\beta}\|^2 + r_n \|\boldsymbol{\beta}_{S_1^c}\|^2\right\}.
$$

Therefore, $\hat{\boldsymbol{\beta}}_n^{\mathrm{WR}} = \arg\min\{L(\boldsymbol{\beta}_n; S_1)\}$ satisfies,

$$\frac{\partial L\left(\hat{\boldsymbol{\beta}}_n^{\mathrm{WR}}\right)}{\partial \boldsymbol{\beta}_{S_3^c}} = \mathbf{0}.$$

From the notation $\mathbf{X}'_{S_3^c} = \mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$. If we write $\mathbf{X}'_{S_3} = (\mathbf{w}_1, \ldots, \mathbf{w}_n)$, then

$$-\sum_{i=1}^{n}\left(y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}_{S_3^c}^{\mathrm{WR}} - \mathbf{w}_i'\hat{\boldsymbol{\beta}}_{S_3}^{\mathrm{WR}}\right)\mathbf{z}_i + r_n\begin{pmatrix}\mathbf{0}_{p_{1n}}\\ \hat{\boldsymbol{\beta}}_{S_2}^{\mathrm{WR}}\end{pmatrix} = \mathbf{0}_{p_{1n}+p_{2n}}.$$

Replacing *yi* by $\mathbf{z}_i'\boldsymbol{\beta}_{0S_3^c} + \mathbf{w}_i'\boldsymbol{\beta}_{0S_3} + \varepsilon_i$, we have

$$-\sum_{i=1}^{n}\left(\varepsilon_i - \mathbf{z}_i'\left(\hat{\boldsymbol{\beta}}_{S_3^c}^{\mathrm{WR}} - \boldsymbol{\beta}_{0S_3^c}\right) - \mathbf{w}_i'\left(\hat{\boldsymbol{\beta}}_{S_3}^{\mathrm{WR}} - \boldsymbol{\beta}_{0S_3}\right)\right)\mathbf{z}_i + r_n\begin{pmatrix}\mathbf{0}_{p_{1n}}\\ \hat{\boldsymbol{\beta}}_{S_2}^{\mathrm{WR}}\end{pmatrix} = \mathbf{0}.$$

Notice that $\boldsymbol{\Sigma}_n = n^{-1}\sum_{i=1}\mathbf{z}_i\mathbf{z}_i' \to \boldsymbol{\Sigma}$. Thus,

$$\begin{aligned} n^{1/2}\mathbf{d}_n'\left(\hat{\boldsymbol{\beta}}_{S_3^c}^{\mathrm{WR}} - \boldsymbol{\beta}_{S_3^c}^{*}\right) = {} & n^{-1/2}\sum_{i=1}^{n}\mathbf{d}_n'\varepsilon_i\boldsymbol{\Sigma}_n^{-1}\mathbf{z}_i - n^{-1/2}r_n\mathbf{d}_n'\boldsymbol{\Sigma}_n^{-1}\begin{pmatrix}\mathbf{0}_{p_{1n}}\\ \hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}\end{pmatrix} \\ & - n^{-1/2}\sum_{i=1}^{n}\mathbf{d}_n'\mathbf{w}_i'\left(\hat{\boldsymbol{\beta}}_{3n}^{\mathrm{WR}} - \boldsymbol{\beta}_{S_3}^{*}\right)\boldsymbol{\Sigma}_n^{-1}\mathbf{z}_i \end{aligned} \qquad \text{(H5)}$$

Under conditions (B1–B3), with probability 1, from Theorem 1. Therefore, the third term in (H5) is zero. By abusing the notation, if we rewrite $\mathbf{d}n = (\mathbf{d}1n', \mathbf{d}2n')'$, then

where the first '≤' is from (B4), the first '=' is from (A2) and (B1), the second '≤' is from the Cauchy–Schwarz inequality, the third '≤' is from (A2). The last '=' holds because $rn = o(n^{1/2-}\tau)$

if we choose with $an = c_1 n^{-}\alpha$ for $\alpha < 1/4 - \tau/2$ for $0 < \tau < 1/2$. Therefore,

$$\lim_{n\to\infty} n^{1/2}s_n^{-1}\mathbf{d}_n'\left(\hat{\boldsymbol{\beta}}_{S_3^c}^{\mathrm{WR}} - \boldsymbol{\beta}_{S_3^c}^{*}\right) = \lim_{n\to\infty} n^{-1/2}s_n^{-1}\sum_{i=1}^{n}\mathbf{d}_n'\varepsilon_i\boldsymbol{\Sigma}_n^{-1}\mathbf{z}_i \qquad \text{(H6)}$$

Define $U_i = n^{-1/2}s_n^{-1}d_n'\Sigma_n^{-1}z_i$, 1 ≤i≤n. From (B1), we know that $\sum_{i=1}^{n}u_i\varepsilon_i$ is normal with variance,

$$\mathrm{Var}\left(\sum_{i=1}^{n}(u_i\varepsilon_i)\right) = \sigma^2 n^{-1}s_n^{-2}\mathbf{d}_n'\boldsymbol{\Sigma}_n^{-1}\left(\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i\right)\boldsymbol{\Sigma}_n^{-1}\mathbf{d}_n = 1.$$

Proof of Theorem 3. First, (4.9a) holds because we have

$$\lim_{n\to\infty} E\left[n^{1/2}s_{1n}^{-1}\mathbf{d}_{1n}'\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}-\boldsymbol{\beta}_1^*\right)\right]^2\Big] = E\left\{\lim_{n\to\infty}\left[n^{1/2}s_{1n}^{-1}\mathbf{d}_{1n}'\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}-\boldsymbol{\beta}_1^*\right)\right]^2\Big]\right\} = E[Z^2] = 1,$$

where $Z \sim N(0,1)$. We now verify (4.9b). Let $\tilde{y} = y - X_{2n}\hat{\beta}_{2n}^{wR} - X_{3n}\hat{\beta}_{3n}$ Then,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} &= \arg\ \min\{\|\tilde{\mathbf{y}} - \mathbf{X}_{1n}\boldsymbol{\beta}_{1n}\|^2\} \\
&= (\mathbf{X}_{1n}'\mathbf{X}_{1n})^{-1}\mathbf{X}_{1n}'\tilde{\mathbf{y}} \\
&= \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}} - (\mathbf{X}_{1n}'\mathbf{X}_{1n})^{-1}\mathbf{X}_{1n}'\mathbf{X}_{2n}\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}} - (\mathbf{X}_{1n}'\mathbf{X}_{1n})^{-1}\mathbf{X}_{1n}'\mathbf{X}_{3n}\hat{\boldsymbol{\beta}}_{3n}^{\mathrm{WR}} \\
&= \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}} - (\mathbf{X}_{1n}'\mathbf{X}_{1n})^{-1}\mathbf{X}_{1n}'\mathbf{X}_{2n}\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}.
\end{aligned} \tag{H7}$$

From the definition,

$$\begin{aligned}
R\left(\mathbf{d}_{1n}'\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right) &= \lim_{n\to\infty} E\left[n^{1/2}s_{1n}^{-1}\mathbf{d}_{1n}'\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}-\boldsymbol{\beta}_1^*\right)\right]^2 \\
&= \lim_{n\to\infty} s_{1n}^{-2}E\left\{n^{1/2}\mathbf{d}_{1n}'\left[\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}-\boldsymbol{\beta}_1^*\right)-\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}-\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right)\right]\right\}^2 \\
&= \lim_{n\to\infty} E\left\{n^{1/2}s_{1n}^{-2}\mathbf{d}_{1n}'\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}-\boldsymbol{\beta}_1^*\right)\right\}^2 + \lim_{n\to\infty} E\left\{n^{1/2}s_{1n}^{-2}\mathbf{d}_{1n}'\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}-\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right)\right\}^2 \\
&\quad - 2\lim_{n\to\infty} E\left\{ns_{1n}^{-2}\mathbf{d}_{1n}'\left(\mathbf{X}_{1n}'\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}-\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right)\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}-\boldsymbol{\beta}_1^*\right)'\mathbf{d}_{1n}\right\} \right. \\
&= I_1 + I_2 + I_3.
\end{aligned}$$

From (4.9a), we know $I_1 = \lim n \to \infty \in \left\{n^{1/2}s_{1n}^{-1}d_1'n\left(\hat{\beta}_{1n}^{wR} - \beta_1^*\right)\right\}^2$. From (H7),

$$\begin{aligned}
I_2 &= \lim_{n\to\infty} s_{1n}^{-2}E\left\{n^{1/2}\mathbf{d}_{1n}'\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}}-\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right)\right\}^2 \\
&= \lim_{n\to\infty} s_{1n}^{-2}E\left\{n^{1/2}\mathbf{d}_{1n}'\boldsymbol{\Sigma}_{n11}^{-1}\boldsymbol{\Sigma}_{n12}\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}\right\}^2 \\
&= \lim_{n\to\infty} \left(s_{2n}^2/s_{1n}^2\right)E\left\{n^{1/2}s_{2n}^{-1}\mathbf{d}_{2n}'\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}\right\}^2,
\end{aligned}$$

where $d_{2n} = \Sigma_{n21}\Sigma_{n11}^{-1}d_1n$ and $S_{2n}^2 = d_{2n}'\Sigma_{n22.1}^{-1}d_{2n}$. From Ouellette (1981) Equation (1.12), we obtain

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22.1}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} = \boldsymbol{\Sigma}_{11.2}^{-1} - \boldsymbol{\Sigma}_{11}^{-1}. \tag{H8}$$

Therefore,

$$s_{2n}^2 = \sigma^2 \mathbf{d}_{1n}' \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22.1}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \mathbf{d}_{1n} = \sigma^2 \mathbf{d}_{1n}' \boldsymbol{\Sigma}_{11.2}^{-1} \mathbf{d}_{1n} - \sigma^2 \mathbf{d}_{1n}' \boldsymbol{\Sigma}_{11}^{-1} \mathbf{d}_{1n}.$$

Because $s_{2n}^2 \,/\, s_{1n}^2 \to 1 - c$,

$$I_2 = (1-c) \lim_{n\to\infty} E\left[\chi_1^2(\Delta_{\mathbf{d}_{1n}})\right] = (1-c)(1 + \Delta_{\mathbf{d}_{1n}}),$$

where $x_v^2(t)$ is a χ^2 distribution with degrees of freedom *v* and noncentral parameter *t*. Here, $\Delta\mathbf{d}_1 n$ is given in (4.8). From the Cauchy–Schwarz inequality,

$$\Delta_{\mathbf{d}_{1n}} = s_{2n}^{-2} (\mathbf{d}_{2n}' \delta)^2 \leqslant \delta' \boldsymbol{\Sigma}_{n22.1} \delta.$$

Furthermore,

$$\begin{aligned} I_3 &= -2 \lim_{n\to\infty} E\left\{ n S_{1n}^{-2} \mathbf{d}_{1n}' \left(\mathbf{X}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}} \right) \left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^* \right)' \mathbf{d}_{1n} \right\} \right. \\ &= 2 \lim_{n\to\infty} \left[\mathbf{d}_{1n}' (\mathbf{I}_{p_{1n}} \; \mathbf{0}_{p_{1n}\times p_{2n}}) \boldsymbol{\Sigma}_n^{-1} (\mathbf{0}'_{p_{1n}\times p_{2n}} \; \mathbf{I}_{p_{2n}})' \boldsymbol{\Sigma}_{n21} \boldsymbol{\Sigma}_{n11}^{-1} \right] \\ &= -2 \lim_{n\to\infty} (s_{2n}/s_{1n})^2 = -2(1-c) \end{aligned}$$

Thus, $R\left(d_{1n}' \hat{\beta}_{1^n}^{RE}\right) = I_1 + I_2 + I_3 = 1 - \left(1 - \Delta_{d_{1n}}\right)(1 - C)$. Thus, (4.9b) holds.

We now investigate (4.9c). First from the definition,

$$\begin{aligned} R\left(\mathbf{d}_{1n}' \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{SE}}\right) &= \lim_{n\to\infty} E[n^{1/2} s_{1n}^{-1} \mathbf{d}_{1n}' (\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{SE}} - \boldsymbol{\beta}_1^*)]^2 \\ &= \lim_{n\to\infty} s_{1n}^{-2} E\left\{ n^{1/2} \mathbf{d}_{1n}' \left[\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^* \right) - (\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}})((p_{2n}-2)/T_n) \right] \right\}^2 \\ &= \lim_{n\to\infty} E\left\{ n^{1/2} s_{1n}^{-2} \mathbf{d}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^* \right) \right\}^2 \\ &\quad - (\lim_{n\to\infty} 2E\left\{ n s_{1n}^{-2} ((p_{2n}-2))/T_n) \mathbf{d}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}} \right) \left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^* \right)' \mathbf{d}_{1n} \right\} \\ &\quad - \lim_{n\to\infty} E\left\{ n^{1/2} s_{1n}^{-2} \mathbf{d}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}} \right) ((p_{2n}-2)/T_n) \right\}^2 \Big) \\ &= J_1 - (J_2 - J_3). \end{aligned}$$

Again, $J_1 = \lim n \to \infty \left\{ n^{1/2} s_{1n}^{-1} d_1' n \left(\hat{\beta}_{1n}^{WR} - \beta_1^* \right) \right\}^2$. From (H7),

$$\mathbf{d}_{1n}' \left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}} \right) = -\mathbf{d}_{1n}' \boldsymbol{\Sigma}_{n11}^{-1} \boldsymbol{\Sigma}_{n12} \hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}} = \mathbf{d}_{2n}' \hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}.$$

Then, we have

$$
\begin{aligned}
J_2 - J_3 = \lim_{n\to\infty} 2E\left\{ ns_{1n}^{-2}((p_{2n}-2)/T_n)\mathbf{d}_{1n}'\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}}\right)\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^*\right)'\mathbf{d}_{1n}\right\} \\
- \lim_{n\to\infty} E\left\{ n^{1/2}s_{1n}^{-2}\mathbf{d}_{1n}'(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \hat{\boldsymbol{\beta}}_{1n}^{\mathrm{RE}})((p_{2n}-2)/T_n)\right\}^2 \\
= -\lim_{n\to\infty} 2s_{1n}^{-2}E\left\{ ((p_{2n}-2)/T_n)\sqrt{n}\mathbf{d}_{2n}'\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{1n}^{\mathrm{WR}} - \boldsymbol{\beta}_1^*\right)'\mathbf{d}_{1n}\right\} \\
- \lim_{n\to\infty} s_{1n}^{-2}E\left\{ \left[((p_{2n}-2)/T_n)\sqrt{n}\mathbf{d}_{2n}'\hat{\boldsymbol{\beta}}_{2n}^{\mathrm{WR}}\right]^2\right\}
\end{aligned}
$$

From Theorem 1, $\hat{s}_2 = p_{2n} + o_p(1)$ and

$$
T_n = \left(\sqrt{n}\hat{\boldsymbol{\beta}}_{n2}^{\mathrm{WR}}\right)'(\boldsymbol{\Sigma}_{n22.1})\left(\sqrt{n}\hat{\boldsymbol{\beta}}_{n2}^{\mathrm{WR}}\right)/\hat{\sigma}^2 + o_p(1).
$$

We now define $\mathbf{a}' = \left(\mathbf{d}_{1n}'\ \mathbf{0}_{1\times p_{2n}}\right), \mathbf{b}' = \left(\mathbf{0}_{p_{1n}\times 1}\ -\mathbf{d}_{2n}\right)$,

and $\eta(\mathbf{x}) = ((p_2 n - 2)/(\mathbf{x}'\mathbf{W}\mathbf{x}))\mathbf{b}'\mathbf{x}$, where $w = \begin{pmatrix} 0 & 0 \\ 0 & \Sigma_{n22.1} \end{pmatrix}$ Then, from the asymptotic normality,

$$
J_2 - J_3 = \lim_{n\to\infty} s_{1n}^{-2}E[2\eta(\mathbf{z}+\boldsymbol{\zeta})\mathbf{a}'\mathbf{z} - (\eta(\mathbf{z}+\boldsymbol{\zeta}))^2],
$$

where $\boldsymbol{\zeta}' = \left(\mathbf{0}_{p_{1n}\times 1}\ -\boldsymbol{\beta}_{20}'\right)$ and $\mathbf{z}$ satisfy that

$$
\left(\mathbf{d}_n'\boldsymbol{\Sigma}_n^{-1/2}\mathbf{d}_n\right)^{-1}\mathbf{d}_n'\mathbf{z} \to N(0,1)
$$

and

$$
\lim_{n\to\infty}\left(\mathbf{d}_n'\boldsymbol{\Sigma}_n^{-1}\mathbf{d}_n\right)^{-1}(\mathbf{d}_n'\boldsymbol{\zeta})^2 = \lim_{n\to\infty}\Delta_{d_{1n}}.
$$

From Stein's lemma, we have

$$E(\eta(\mathbf{z}+\boldsymbol{\zeta})\mathbf{a}'\mathbf{z}) = \mathbf{a}'\boldsymbol{\Sigma}_n^{-1}(\partial\eta(\mathbf{z}+\boldsymbol{\zeta})/\partial\mathbf{z})$$
$$= \frac{(p_{2n}-2)\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{b}}{(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})} - \frac{2(p_{2n}-2)\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{b}}{((\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta}))^2}$$

So we have

$$J_2 - J_3 = \lim_{n\to\infty} s_{1n}^{-2}E[2\eta(\mathbf{z}+\boldsymbol{\zeta})\mathbf{a}'\mathbf{z} - (\eta(\mathbf{z}+\boldsymbol{\zeta}))^2]$$
$$= \lim_{n\to\infty} s_{1n}^{-2}E\left\{\left[2\frac{(p_{2n}-2)\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{b}}{(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})} - 4\frac{(p_{2n}-2)\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{b}}{((\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta}))^2}\right]\right\}$$
$$- \lim_{n\to\infty} s_{1n}^{-2}E\left\{\frac{(p_{2n}-2)^2\mathbf{b}'(\mathbf{z}+\boldsymbol{\zeta})(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{b}}{((\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta}))^2}\right\}$$
$$= \lim_{n\to\infty} E\left\{\frac{(p_{2n}-2)}{(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}f\right\},$$

where

$$f = \frac{2\mathbf{a}'\boldsymbol{\Sigma}_n^{-1}\mathbf{b}}{s_{1n}^2} - \frac{4(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}\boldsymbol{\Sigma}_n^{-1}\mathbf{a}\mathbf{b}'(\mathbf{z}+\boldsymbol{\zeta})}{s_{1n}^2(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})} - \frac{(p_{2n}-2)(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{b}\mathbf{b}'(\mathbf{z}+\boldsymbol{\zeta})}{s_{1n}^2(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}.$$

Notice that $a'\Sigma_n^{-1}b = d_{2n}'\Sigma_{n21.1}^{-1}d_{2n} = s_{2n}^2$ and $W\Sigma_n^{-1}ab' = bb'$ Therefore,

$$f = 2\frac{s_{2n}^2}{s_{1n}^2} - \frac{(p_{2n}+2)(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{b}\mathbf{b}'(\mathbf{z}+\boldsymbol{\zeta})}{s_{1n}^2(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}.$$

Thus,

$$J_2 - J_3 = \lim_{n\to\infty} E\left\{\frac{(p_{2n}-2)}{(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}\left[\frac{2s_{2n}^2}{s_{1n}^2} - \frac{(p_{2n}+2)(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{d}_{2n}\mathbf{d}_{2n}'(\mathbf{z}+\boldsymbol{\zeta})}{s_{1n}^2(\mathbf{z}+\boldsymbol{\zeta})'\mathbf{W}(\mathbf{z}+\boldsymbol{\zeta})}\right]\right\}$$
$$= \lim_{n\to\infty} \frac{s_{2n}^2}{s_{1n}^2}E\left\{\frac{(p_{2n}-2)}{(\mathbf{z}_2+\boldsymbol{\delta})'\boldsymbol{\Sigma}_{n22.1}(\mathbf{z}_2+\boldsymbol{\delta})}\left[2 - \frac{(p_{2n}+2)(\mathbf{z}_2+\boldsymbol{\delta})'\mathbf{d}_{2n}\mathbf{d}_{2n}'(\mathbf{z}_2+\boldsymbol{\delta})}{s_{2n}^2(\mathbf{z}_2+\boldsymbol{\delta})'\boldsymbol{\Sigma}_{n22.1}(\mathbf{z}_2+\boldsymbol{\delta})}\right]\right\},$$

where $\mathbf{z}_2$ satisfies that $s_{2n}^{-1}d_{2n}'z_2 \to N(0,1)$. Thus (4.9c) holds. Similarly, we can obtain (4.9d).

Proof of Corollary 3. We first verify (i).

Define $\tilde{\mathbf{z}}_2 = \sigma^{-2}\boldsymbol{\Sigma}_{n22.1}^{1/2}(\mathbf{z}_2 + \boldsymbol{\delta})$ and $\mathbf{B} = (\sigma^2/s_{2n}^2)\boldsymbol{\Sigma}_{n22.1}^{-1/2}\mathbf{d}_{2n}\mathbf{d}_{2n}'\boldsymbol{\Sigma}_{n22.1}^{-1/2}$. From the Cramér–Wold device, we have

$$\begin{aligned} J_2 - J_3 &= (1-c)\lim_{n\to\infty}\left\{E\left[\frac{2(p_{2n}-2)}{\tilde{\mathbf{z}}_2'\tilde{\mathbf{z}}_2}\right] - E\left[\frac{(p_{2n}-2)(p_{2n}+2)\tilde{\mathbf{z}}_2'\mathbf{B}\tilde{\mathbf{z}}_2}{(\tilde{\mathbf{z}}_2'\tilde{\mathbf{z}}_2)^2}\right]\right\} \\ &= (1-c)\lim_{n\to\infty}\left\{E\left[\frac{p_{2n}-2}{\chi^2_{p_{2n}}(\Delta_n)}\right] - \mathrm{Tr}(\mathbf{B})E\left[\frac{(p_{2n}-2)(p_{2n}+2)}{\chi^4_{p_{2n}+2}(\Delta_n)}\right]\right\} \\ &\quad + (1-c)\lim_{n\to\infty}\left\{E\left[\frac{p_{2n}-2}{\chi^2_{p_{2n}}(\Delta_n)}\right] - (\boldsymbol{\delta}'\mathbf{B}\boldsymbol{\delta})E\left[\frac{(p_{2n}-2)(p_{2n}+2)}{\chi^2_{p_{2n}+4}(\Delta_n)}\right]\right\}, \end{aligned}$$

where $\Delta n = \boldsymbol{\delta}'\boldsymbol{\Sigma}n_{22.1}\boldsymbol{\delta}$ and 'Tr(**B**)' is the trace of matrix **B**. Here, the second '=' is from Theorem 8 in Chapter 2 in [29]. Notice that Tr(**B**) = 1. Using the relationship between the chi-square distribution and Poisson distribution,

$$\begin{aligned} J_2 - J_3 &= (1-c)\lim_{n\to\infty}\left\{E_\kappa\left[\frac{p_{2n}-2}{p_{2n}-2+2\kappa}\left(1-\frac{p_{2n}+2}{p_{2n}+2\kappa}\right)\right]\right\} \\ &\quad + (1-c)\lim_{n\to\infty}\left\{E_\kappa\left[\frac{p_{2n}-2}{p_{2n}-2+2\kappa}\left(1-\frac{\boldsymbol{\delta}'\mathbf{B}\boldsymbol{\delta}(p_{2n}-2+2\kappa)(p_{2n}+2)}{(p_{2n}+2+2\kappa)(p_{2n}+2\kappa)}\right)\right]\right\}, \end{aligned}$$

where κ is a Poisson distribution with mean $\Delta n/2$ and $E\kappa$ means the expectation is taken for the Poisson random variable κ. Because $P(k \geq 1)$ when $p_2 n \to \infty$. With almost probability 1, we have

$$0 \leqslant \frac{p_{2n}-2}{p_{2n}-2+2\kappa}\left(1-\frac{p_{2n}+2}{p_{2n}+2\kappa}\right) \leqslant 1.$$

If $\|\boldsymbol{\delta}\|^2 \leq 1$, then $\boldsymbol{\delta}'\mathbf{B}\boldsymbol{\delta} = \left(\boldsymbol{\delta}'\boldsymbol{\Sigma}_{n22.1}^{-1/2}\mathbf{d}_{2n}\right)^2 / \left(\mathbf{d}_{2n}'\boldsymbol{\Sigma}_{n22.1}^{-1}\mathbf{d}_{2n}\right) \leqslant \boldsymbol{\delta}'\boldsymbol{\delta} \leqslant 1.$. Then, $E[g_1(z_2 + \delta]$. Furthermore, when $\mathbf{x}'\boldsymbol{\Sigma}n_{22.1}\mathbf{x} \leq p_2 n - 2$, we have

$$2 - s_{2n}^{-2}\mathbf{x}\mathbf{d}_{2n}\mathbf{d}_{2n}'\mathbf{x}' \geqslant 2 - \frac{(p_{2n}-2)\mathbf{x}\mathbf{d}_{2n}\mathbf{d}_{2n}'\mathbf{x}'}{s_{2n}^2\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}} > 2 - \frac{(p_{2n}+2)\mathbf{x}\mathbf{d}_{2n}\mathbf{d}_{2n}'\mathbf{x}'}{s_{2n}^2\mathbf{x}'\boldsymbol{\Sigma}_{n22.1}\mathbf{x}}.$$

Therefore, $g_2(x) \geq g_1(x)$. Thus, (i) holds.

In fact, the inequalities in (i) also hold even though $\|\boldsymbol{\delta}\|^2 > 1$. For example, suppose $\Delta n = \iota p_2 n$ for some constant $\iota > 0$. Then, $p_{2n}^{-1/2}(2\kappa - \Delta_n) \rightsquigarrow N(0, \iota^{-1})$. Therefore, if $\|\boldsymbol{\delta}\|^2 \leq 1 + \iota$, with probability 1, we have

$$1 - \frac{\boldsymbol{\delta}'\mathbf{B}\boldsymbol{\delta}(p_{2n} - 2 + 2\kappa)(p_{2n} + 2)}{(p_{2n} + 2 + 2\kappa)(p_{2n} + 2\kappa)} \to 1 - \frac{\|\boldsymbol{\delta}\|^2}{1 + \iota} > 0.$$

Thus, (ii) holds.

We now verify (iii). If $\boldsymbol{\delta} = \mathbf{0}$, then $\Delta \mathbf{d}_1 n = 0$. Thus, $\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n}) = c < \mathrm{ADR}(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{WR}}_{1n})$. We now compare $\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{SE}}_{1n})$ with $\mathrm{ADR}(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n})$.

Denote $\mathbf{A} = (p_{2n} + 2)s^2_{2n}\boldsymbol{\Sigma}^{-1/2}_{n22.1}\mathbf{d}_{2n}\mathbf{d}'_{2n}\boldsymbol{\Sigma}^{-1/2}_{n22.1}$. If $\boldsymbol{\delta} = \mathbf{0}$, then we have

$$(1 - c)^{-1} g_1(\mathbf{z}_2) = \lim_{n\to\infty} \frac{(p_{2n} - 2)}{\mathbf{z}'_2\boldsymbol{\Sigma}_{n22.1}\mathbf{z}_2} - \lim_{n\to\infty} \frac{(p_{2n} - 2)\left(\boldsymbol{\Sigma}^{1/2}_{n22.1}\mathbf{z}_2\right)' \mathbf{A} \left(\boldsymbol{\Sigma}^{1/2}_{n22.1}\mathbf{z}_2\right)}{\left(\mathbf{z}'_2\boldsymbol{\Sigma}_{n22.1}\mathbf{z}_2\right)^2}.$$

From Theorem 2.1.8 in [29] and moment of inverse chi-squares distribution, we have

$$\lim_{n\to\infty} E\left[\frac{(p_{2n} - 2)}{\mathbf{z}'_2\boldsymbol{\Sigma}_{n22.1}\mathbf{z}_2}\right] = 2$$

and

$$\lim_{n\to\infty} E\left[\frac{(p_{2n} - 2)\left(\boldsymbol{\Sigma}^{1/2}_{n22.1}\mathbf{z}_2\right)' \mathbf{A} \left(\boldsymbol{\Sigma}^{1/2}_{n22.1}\mathbf{z}_2\right)}{\left(\mathbf{z}'_2\boldsymbol{\Sigma}_{n22.1}\mathbf{z}_2\right)^2}\right] = \lim_{n\to\infty} 1 + 2/p_{2n}.$$

Thus, if $p_2 n = p_2$ is fixed, $E[g_1(\mathbf{z}_2)] = (1 - c)(1 - 2/p_2) < 1 - c$. Therefore,

$$\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{SE}}_{1n}\right) > \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n}\right).$$

Similarly, we can verify $\mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{RE}}_{1n}\right) < \mathrm{ADR}\left(\mathbf{d}'_{1n}\hat{\boldsymbol{\beta}}^{\mathrm{PSE}}_{1n}\right)$ When $p_2 n \to \infty$, .

References

1 Tibshirani R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 1996; **58**:267–288.

2 Leng C, Lin Y, Wahba G. A note on the Lasso and related procedures in model selection. *Statistica Sinica* 2006; **16**:1273–1284.

3 Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001; **96**:1348–1360.

4 Fan J, Lv J. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory* 2011; **57**(8):5467–5484.

5 Zou H. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* 2006; **101**:1418–1429.

6 Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 2010; **38**(2):894–942.

7 Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 2010; **20**(1):101.

8 Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B* 2014; **76**(1):217–242.

9 Zhao P, Yu B. On model selection consistency of LASSO. *Journal of Machine Learning Research* 2006; **7**:2541–2563.

10 Huang J, Ma SG, Zhang CH. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica* 2008; **18**:1603–1618.

11 Bickel P, Ritov Y, Tsybakov A. Simultaneous analysis of Lasso and dantzig selector. *Annals of Statistics* 2009; **37**:1705–1732.

12 Hansen BE. The risk of James-Stein and Lasso shrinkage, 2013. http://www.ssc.wisc.edu/bhansen/papers/lasso.pdf [accessed 20 December 2015] .

13 Belloni A, Chernozhukov V. Least squares after model selection in high-dimensional sparse models. *Bernoulli* 2009; **19**:521–547.

14 Liu H, Yu B. Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics* 2013; **7**:3124–3169.

15 Stein C. 1956. Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955*, vol. **I**. University of California Press: Berkeley and Los Angeles; 187–195.

16 James W, Stein C. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1961; **1**:361–379.

17 Ahmed SE. *Penalty, Shrinkage and Pretest Strategies: Variable Selection and Estimation*. Springer: New York, 2014.

18 Ahmed SE, Fallahpour S. Shrinkage estimation strategy in quasi-likelihood models. *Statistics & Probability Letters* 2012; **82**:2170?2179.

19 Ahmed SE, Hossain S, Doksum KA. Lasso and shrinkage estimation in Weibull censored regression models. *Journal of Statistical Inference and Planning* 2012; **12**:1273–1284.

20 Ahmed SE, Doksum KA, Hossain S, You J. Shrinkage, pretest and absolute penalty estimators in partially linear models. *Australian & New Zealand Journal of Statistics* 2007; **49**:435–454.

21 Marsaglia G, Styan GPH. Equalities and inequalities for ranks of matrices. *Linear and Multilinear Algebra* 1974; **2**:269–292.

22 Frank IE, Friedman JH. A statistical view of some chemometrics regression tools (with discussion). *Technometrics* 1993; **35**:109–148.

23 Zhang CH, Huang J. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *The Annals of Statistics*2008; **36**:1567–1594.

24 Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* 1998; **20**(1):3361.

25 Wainwright MJ. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* 2009; **55**(5):2183–2202.

26 Zheng Z, Fan Y, Lv J. High dimensional threshold regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B*2014; **76**:627–649.

27 Weng H, Feng Y, Qiao X. Regularization after retention in ultrahigh dimensional linear regression models, 2013. arXiv preprint arXiv:1311.5625.

28 Shao J, Deng X. Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics* 2012; **40**:812–831.

29 Saleh A. KME. *Theory of Preliminary Test and Stein-Type Estimation with Applications*. Wiley: New York, 2006.

30 Barro R, Lee J. Data set for a panel of 139 countries, 1994. http://admin.nber.org/pub/barro.lee/ [accessed 20 December 2015].

31 Barro R, Sala-i Martin X. *Economic Growth*. McGraw-Hill: New York, 1995.

32 Durlauf S, Johnson PA, Temple JRW. Growth econometrics. *Handbook of Economic Growth* 2005; **1**:555–677.

33 Lee S, Seo MH, Y S. The LASSO for high-dimensional regression with a possible change-poin. *Journal of the Royal Statistical Society: Series B* 2016; **78**:193–210.